8-21-2020

Chemical Tools and Mass Spectrometry-based Approaches for Exploring Reactivity and Selectivity of Small Molecules in Complex Proteomes

Lei Wang
University of Connecticut - Storrs, lei.6.wang@uconn.edu

Chemical Tools and Mass Spectrometry-based Approaches for Exploring Reactivity and Selectivity of Small Molecules in Complex Proteomes

Lei Wang, PhD

University of Connecticut, 2020

Abstract

The biomedical and pharmaceutical communities are experiencing a growing demand for new druggable targets. On the other hand, many marketed drugs are being repurposed for their newly uncovered pharmacological activities. In fact, off-target drug effects can either lead to adverse events or new marketable indications. As the healthcare industry moves closer toward systems biology and precision medicine, comprehensive profiling of small molecule-protein interactions has become increasingly crucial. Chemical proteomics is a powerful set of bioanalytical approaches that utilize small molecule chemical probes and affinity capture mass spectrometry to study proteome-wide actions of reactive small molecules like drugs, toxins, and metabolites.

This dissertation discusses multiple technical and methodological aspects of chemical proteomics as a multidisciplinary subject. It presents two projects that exemplify the development of chemical probes and implementation of chemical proteomics in two distinct directions known as compound-centric and activity-based. Within both studies, the modification-specific data processing principle has provoked awareness and thoughts on tailoring bioinformatics tools for chemical proteomics.

The 2-nitroimidazole-indocyanine green (2-nitro-ICG) project features the deployment of a novel compound-centric chemical probe that answers how 2-nitroimidazole targets tumor hypoxia. This study concludes that 2-nitro-ICG and its reduced fragments modify mouse albumin as the primary target, but at two distinct sites via two different mechanisms. The development and application of 2-nitro-ICG also demonstrate various analytical benefits, challenges, and pitfalls in the compound-centric direction of chemical proteomics.

The α-methylene-β-lactone (MeLac) project presents an innovative activity-based probe with multiple electrophilic sites. This study explores the significance of broad probe reactivity in activity-based chemical proteomics. It concludes that MeLac is reactive to amino, hydroxyl, and thiol groups on proteins. Moreover, the discovery of MeLac-alkyne glutathione adduct reveals a potential shortcut for customizing compound-centric probes. Therefore, the multi-electrophilic MeLac moiety creates both a versatile activity-based probe and a scaffold for assembling compound-centric probes.

Overall, chemical proteomics has established its irreplaceable position in chemical biology. Researchers exert efforts to make chemical proteomics technologies more applicable. The chemical tools, analytical workflows, and bioinformatics support continue to improve. The field of chemical proteomics will expand faster as more innovations emerge at its cutting edge.

Chemical Tools and Mass Spectrometry-based Approaches for Exploring Reactivity and Selectivity of Small Molecules in Complex Proteomes

Lei Wang

B.S., Iowa State University, 2012

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020

APPROVAL PAGE

Doctor of Philosophy Dissertation

Chemical Tools and Mass Spectrometry-based Approaches for Exploring Reactivity and Selectivity of Small Molecules in Complex Proteomes

Presented by

Lei Wang, B.S.

Major Advisor _____
Xudong Yao, Ph.D.

Associate Advisor _____
Alfredo Angeles-Boza, Ph.D.

Associate Advisor _____
Christian Brückner, Ph.D.

Associate Advisor _____
Mark W. Peczuh, Ph.D.

Associate Advisor _____
Fatma Selampinar, Ph.D.

University of Connecticut

2020

Acknowledgments

First of all, I would like to express my deepest appreciation and gratitude to my major advisor, Dr. Xudong Yao, for his enlightening guidance and impeccable mentorship. He provided guidance patiently and warmly throughout my PhD journey at and beyond UConn Chemistry Department. With immense intelligence, wisdom, and resourcefulness, Dr. Yao never fails to help me tackle overwhelming challenges, provide respite from hardship, and guide me through career milestones. I am truly fortunate to have had the opportunity to become his student.

Second, I would like to thank my committee members, Drs. Alfredo Angeles-Boza, Christian Brückner, Mark Peczuh, and Fatma Selampinar for their supportive guidance, thought-provoking suggestions, and academic collegiality over the years. Similarly, I would like to recognize my senior lab mentors, Dr. Bekim Bajrami, for his tremendous help and mentorship during my internship, and Drs. Adam McShane, Vahid Farrokhi, Mary Joan Castillo, Reza Nemati, and Yuanyuan Shen for the contributions that each of them made to my intellectual growth during my years of study. My gratitude also goes to my former and current lab mates, Song Li, Veronica Cheng, Louis Riel Pascal, Mariel Clores, Clodette Punzalan, and Huidi Tian, for their academic discussions and help with lab routines and my research.

Third, I would like to acknowledge my collaborators, Drs. Amy Howell, Michael Smith, Christopher Dietz, Quing Zhu, Feifei Zhou, and Mohsen Erfanzadeh, who made my dissertation projects possible; my mentors in various mass spectrometry facilities at UConn, Drs. Youjun Fu, Adam Graichen, James Stuart, Jeremy Balsbaugh, and Anthony Provatas for their helpful advisement and research assistance.

Finally, I feel obligated to acknowledge my best friends at UConn and family: my parents for the emotional and financial support for my study and life, my girlfriend Dr. Yangzhou Li for

her constant companionship, and Prof. Jeannette Wick for her courtesy and parent-like hospitality.

I would like to dedicate this work to all my beloved people here.

Thank you!

Lei Wang

Storrs, Connecticut

July 2020

Table of Content

List of Abbreviations and Acronyms

2D	two-dimensional
2-DE	two-dimensional gel electrophoresis
2-nitro-ICG	2-nitroimidazole-indocyanine green
3D	three-dimensional
ABP	activity-based probe
ABPP	activity-based protein profiling
ADME	absorption, distribution, metabolism, and excretion
ADME-Tox	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AGC	automatic gain control
AIF	all-ion-fragmentation
AlbM	albumin
BCA	bicinchoninic acid
C18	octadecyl
CAD	collisional activation dissociation
cDNA	complementary DNA
CID	collision-induced dissociation
cm	centimeter
Co-IP	complex immunoprecipitation
C-terminus	carboxyl-terminus
CuAAC	copper-catalyzed alkyne-azide cycloaddition
Da	Dalton
DDA	data-dependent acquisition
Dde	Dde biotin picolyl azide
Des	desthiobiotin azide
DFT	density functional theory
DIA	data-independent acquisition
Dia	diazo biotin azide
DMAC	dimethylacetamide
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DTE	dithioerythritol
ECD	electron capture dissociation
ER+	estrogen receptor-positive
ESI	electrospray ionization
ETD	electron-transfer dissociation
FA	formic acid
FDR	false discovery rate
FT-ICR	Fourier-transform ion cyclotron resonance

GC-MS	gas chromatography-mass spectrometry
GSH	glutathione
GSH-Lac	α-methylene-β-lactone alkyne glutathione adduct
GSTP1	glutathione S-transferase P1
h	hour
HCD	higher-energy collisional dissociation
HILIC	hydrophilic interaction chromatography
HpH	high-pH reversed-phase fractionation
HPLC	high-performance liquid chromatography
HR/AM	high-resolving power/accurate-mass
IAA	iodoacetamide
ICG	indocyanine green
IEF-PAGE	isoelectric focusing-polyacrylamide gel electrophoresis
IgG	immunoglobulin G
IP	immunoprecipitation
IRC	intrinsic reaction coordinate
Kd	dissociation constant
kV	kilovolt
LC	liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
LC-MS/MS	liquid chromatography-tandem mass spectrometry
LFQ	label-free quantitation
LOQ	limit of quantitation
M	molar
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption/ionization
MeLac	α-methylene-β-lactone
mg	miligram
mg/mL	miligram per mililiter
min	minute
mL	mililiter
mM	milimolar
MODa	modification via alignment
MRM	multiple reaction monitoring
mRNA	messenger ribonucleic acid
MS	mass spectrometry or mass spectrum
MS/MS	tandem mass spectrometry or mass spectrum
MS1	precursor mass spectrum
MS2	two-stage tandem mass spectrometry or mass spectrum
MSE	simultaneous acquisition of exact mass at high and low collision energy
MS^n	multi-stage tandem mass spectrometry or mass spectrum

MSX-DIA	multiplexed data-independent acquisition
MudPIT	multidimensional protein identification technology
NCE	normalized collision energy
NIR	near-infrared
nm	nanometer
NP-40	octyl phenoxypolyethoxylethanol
NSI	nanospray ionization
N-terminus	amino-terminus
NTs	nucleotides
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
PCM	polarizable continuum model
PD	pharmacodynamics
PDB	protein data bank
PET	positron emission tomography
PK	pharmacokinetics
ppm	part per million
PRM	parallel reaction monitoring
PSM	peptide-spectrum matching
PTM	post-translational modification
PVDF	polyvinylidene fluoride
Q	quadrupole mass analyzer
Q1	first quadrupole as the first mass analyzer
q2	second quadrupole as the collision cell
Q3	third quadrupole as the second mass analyzer
Q-Orbi	quadrupole-orbitrap tandem mass spectrometer
QqQ	triple quadrupole tandem mass spectrometer
rcf	relative centrifugal force
SCRF	self-consistent reaction field
SCX	strong cation exchange
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
SILAC	stable isotope by amino acids in cell culture
SILAM	stable isotope labeling by amino acids in mammals
SPE	solid-phase extraction
SRM	selected reaction monitoring
ssDNA	single-stranded deoxyribonucleic acid
SWATH	sequential windowed acquisition of all theoretical fragment ion mass spectra
TAMRA	tetramethylrhodamine
TBS	tris(hydroxymethyl)aminomethane-buffered saline

TBTA	tris(benzyltriazolylmethyl)amine
TCEP	tris(2-carboxyethyl)phosphine
TEAB	triethylammonium bicarbonate
TIC	total ion chromatogram
TMT	tandem mass tag
TOF	time-of-flight mass analyzer
Tris-HCl	tris(hydroxymethyl)aminomethane hydrochloride
TS	transition state
UHPLC	ultra-high-performance liquid chromatography
uMRM	ultra-throughput multiple reaction monitoring
UV	ultraviolet
UVPD	ultraviolet photodissociation
V	volt
vIEF-PAGE	vertical isoelectric focusing-polyacrylamide gel electrophoresis
xg	times gravity
XIC	extracted ion chromatogram
XP	extra precision
μL	microliter
μM	micromolar

Chapter 1 Introduction to Chemical Proteomics

This chapter begins with a briefly introduction to the omics concept and its application in modern biomedical and pharmaceutical research. Afterwards, it mainly focuses on a comprehensive overview of bottom-up proteomics. The discussion touches multiple technical aspects on sample preparation, analytical instrumentation, chemical tools, analytical methodologies, and bioinformatics throughout a typical mass spectrometry-based bottom-up proteomics workflow. Finally, the discussion extends to the introduction to chemical proteomics involving application of these chemical tools and analytical approaches for studying proteome-wide activity of small molecule compounds. The final discussion covers several challenges and their related prospects in the field of chemical proteomics.

1.1 Multi-Omics as a Promising Path to Precision Healthcare

The universal detection and measurement of whole sets of biomolecules shape the technological fundamental of omics research. The principle that a complex biological system can be understood better if treated as a whole defines the characteristic methodology of the omics approach.

Modern biomedical and pharmaceutical research often generates and tests hypotheses by exploring the context of organelles, cells, tissues, organs, and organisms at the molecular level. These explorations scrutinize functions and compositions of individual biological molecules and address their interactions, relationships, and combined influence on an organism's growth and development. The omics approach is particularly suitable for those explorations. In hypothesis-generating experiments, the omics approach acquires and analyzes all available data to construct a hypothesis. In hypothesis-testing experiments, the omics approach highlights nuance, models trends, and builds connections among various subsets of multiple complex biological samples. It often discovers missing pieces from the existing knowledgebase. Moreover, the holistic datasets accumulated in omics research are ready for retrograde analyses if necessary, even for completely irrelevant objectives.

Thus, omics research naturally produces archivable data with great long-term values. A complete biological sample only provides four sets of biological molecules, DNA/genes (genome), mRNA (transcriptome), proteins (proteome), and metabolites (metabolome). From the genome to metabolome, the volume of information expands as molecular diversity increases. (**Figure 1.1**) Given the biological and chemical differences and similarities among these molecules, omics research uses a variety of bioanalytical strategies and techniques. While genomics and transcriptomics primarily rely on high-throughput sequencing technology using fluorescence

detection, proteomics and metabolomics employ liquid chromatography-tandem mass spectrometry (LC-MS/MS) as the main analytical platform.

The principle of omics dates to the early 2000s with the completion of the Human Genome Project.[1] The establishment of automated high-throughput DNA sequencing and synthesis pipelines formed the technological foundation for modern genomics research.[2,3] Meanwhile, the introduction of complementary DNA (cDNA) microarray and mass spectrometry-based protein profiling technologies also enabled the quantitative analysis of the differential global gene expression at messenger RNA level and protein level.[4] These powerful tools marked a historical milestone in transcriptomics and proteomics research. On the other hand, the growing availability of spectroscopy, mass spectrometry, and nuclear magnetic resonance instruments allowed functional analysis at the metabolite level, which was initially referred to as metabolic profiling and metabolic control analysis.[5] It was remarkable that the concept of complete multi-omics study on the yeast model was reported as early as 1998.[5,6]

Nowadays, proteomics and metabolomics research frequently contribute to the development of biomolecular indicators known as biomarkers[7,8] for disease diagnosis[9], progression monitoring[10], drug absorption, distribution, metabolism, and excretion (ADME) in pharmacokinetics and pharmacology (PK/PD)[11] studies. (**Figure 1.2**) Moreover, proteomics platforms are also implemented with chemical tools that facilitate the analyses of intermolecular interactions of proteins with other molecules. For instance, covalent chemical probes that react and label proteins can analyze small molecule drug candidates' effects on the target protein activity in various stages of drug development, such as evaluating drug-target engagement, revealing off-target effect, and assessing drug toxicity.[12] Protein cross-linking reagents that react and conjugate proximate protein pairs and assemblies can analyze protein-protein interactions and reveal

biological roles of protein complexes.[13] Complementary to genomics and proteomics platforms, advanced metabolomics platforms rapidly expand knowledgebase of biochemical pathways and metabolic networks in various organisms, which has led to recent achievements in studies on gut microbiome-related regulations of human immune system.[14]

Analytically, omics samples are essentially mixtures of diverse biomolecules that require pre-analysis preparation. Depending on the specific analytical goal, separation/extraction techniques, which preserve a subgroup of these biomolecules as analytes of interest and remove the rest as interference, are practiced accordingly. The protein or metabolite analytes in the prepared samples are then resolved and measured at molecular level on liquid or gas chromatography-mass spectrometry (LC-MS or GC-MS) systems. While high-resolving power/accurate-mass (HR/AM) mass spectrometers enable unambiguous measurements of biomolecules, ultra-high-performance liquid chromatography (UHPLC) systems provide extra separation of various analytes online. Advanced mass spectrometers are often equipped with instrumental components that can break down ions of analytes into fragments for tandem mass spectrometry (MS/MS) measurements. These fragments contain more useful information as molecular fingerprints. The capability of generating and measuring gas-phase fragments further reinforces the analytical impact of mass spectrometers on the omics research. In fact, researchers routinely use selective measurement of fragment ions generated in the gas phase as the method of choice for quantifying biomolecules of interest in quantitative proteomics and metabolomics. Apparently, the LC-MS/MS, a hyphenated analytical technology, has become the technological core of the proteomics and metabolomics research.

Overall, contemporary multi-omics platforms (genomics, transcriptomics, proteomics, and metabolomics) together with advanced bioinformatics tools and high-performance computing

systems are driving the fast advancement of multidisciplinary biomedical subjects such as systems biology, synthetic biotics[8] and translational medicine[15], which ultimately enact precision and personalized medicine as the state-of-the-art approach to healthcare.

1.2 The Bottom-up Proteomics Workflow

1.2.1 Overview of proteomics

Proteomics is the comprehensive study of proteins. It deciphers protein structures, post-translational modifications (PTMs), cellular functions, biological/pathological significances, and interactions with other molecules. Since proteins are the most important cellular machinery and building blocks, proteomics investigations are being conducted at an escalating rate to explore pathologies of a wide range of diseases and medical conditions including but not limited to cancer[7,16,17], neuro-degenerative disease[18,19], infectious disease[20,21], metabolic disorders[22,23], autoimmune disease[24,25], and cardiovascular disease[26]. Proteomics studies routinely involve the large-scale experimental analysis of proteins.

Depending on the specific study objective, proteins are analyzed either in their original forms (native, slightly reduced, or denatured) or proteolytically cleaved forms. This analytical choice further splits proteomics into two strategic divisions as top-down proteomics and bottom-up proteomics (also known as shotgun proteomics[27]). Each of them has its unique sets of experimental practices and analytical emphases. In general, the top-down strategy is more suitable for protein structure elucidation of purified protein samples. It has the potential to access complete protein sequences and ability to locate and characterize PTMs. It also eliminates the protein digestion step that is most time-consuming but mandatory part of the sample preparation for bottom-up proteomics. However, top-down proteomics is a relatively young field.[28] It requires the most advanced instrumentation and higher operation costs while suffering from a large number of analytical limitations. In comparison to the top-down strategy, the bottom-up strategy is the method of choice for analyzing complex protein samples with higher reliability and lower costs.

Nevertheless, all proteomics experiments often require constant development and implementation of effective analytical techniques. The present collection of various technical disciplines that contributes to proteomics include gel electrophoresis, chromatography, affinity pull-down/immunoprecipitation, spectrophotometry, fluorophotometry, fluorescence imaging, immunoblotting, and most importantly mass spectrometry. In general, a typical bottom-up proteomics workflow (**Figure 1.3**) consists of three major steps as sample preparation, instrumental sample analysis, and data processing for biological inferences.

1.2.2 Sample preparation

The proteomics workflow begins with preparation of complex biological samples. The overall purpose of proteomics sample preparation is to collect analytes of interest from the samples and convert them into instrument-compatible forms. The implementation of a specific sample preparation workflow entails a balance of method efficiency, reproducibility, and cost. A typical proteomics sample preparation workflow is a four-step procedure that involves a series of physical and chemical processes. These four steps are cell lysis, protein separation/clean-up, proteolytic cleavage/digestion, and peptide separation/clean-up.

1.2.2.1 Cell lysis

As the first step, crude biological samples, such as cell culture pellets, animal tissues, and organs, are physically homogenized in a lysis buffer that consists of buffering agents, protease inhibitors, and a surfactant such as octyl phenoxypolyethoxylethanol (NP-40) or Triton X-100 at refrigeration temperature. Although homogenization can be a manual process that takes place in a Dounce tissue grinder, advanced mechanical forces introduced by focused bursts of ultrasonic waves or high-frequency bead beating are usually preferred for extraction of whole-cell protein contents. The lysis process disturbs the cellular membrane while preserving and solubilizing

cellular protein contents. Particulates are usually removed from the samples either via high-speed centrifugation or filtration through 0.2 μm filters. Additional cell fractionation is necessary for more efficient extraction of low-abundance proteins. Mild surfactants such as saponin and digitonin are suitable for the extraction of cytosolic proteins. In contrast, strong ionic surfactants such as sodium dodecyl sulfate and sodium deoxycholate are required for the extraction of nuclear proteins. However, these ionic surfactants can denature proteins and are incompatible with experiments that require protein integrity.

For some special human or animal proteome samples such as plasma and sera, the initial preparation steps generally focus on the depletion of high-abundant proteins rather than solubilization of the whole proteome. Plasma samples consist of mainly high-abundance proteins such as albumin, haptoglobin, hemopexin, immunoglobulins, which are usually not the target proteins of interest. Large quantities of immobilized antibodies on agarose resins are usually used to scavenge these high-abundance inference proteins. These immunodepleting materials are commercially available at relatively high costs.

1.2.2.2 Total protein quantitation assay

It is crucial to measure the total protein concentration of lysates. Although the concentration of a protein solution can be directly determined from spectrophotometric measurements at 280 nm for its UV absorption, reagent and dye-based assays are preferred for higher analytical performance. Several conventional protein assay techniques are available. The bicinchoninic acid (BCA) assay[29] and Lowry assay[30] are redox reaction-based assays, where the protein reduces copper(II) ions to copper(I) ions that form a colored complex. It strongly absorbs light at a specific wavelength (562 nm for BCA assays and 660 nm for Lowry assays). Alternatively, Bradford protein assay[31] is a colorimetric dye-based assay, where the Coomassie

Blue G-250 dye undergoes a light absorption spectrum shift upon binding to proteins. Therefore, the absorbance at 595 nm is proportional to the amount of protein-bound dye, and thus to the sample's protein concentration. Reagents for these assays are commercialized and used routinely in proteomics research. In a total protein quantification assay experiment, reagents are mixed with protein solutions to produce a measurable color change in proportion to the protein amount. The total protein concentration of a lysate is determined by referring to a calibration curve constructed with the assay readout of several known concentrations of a purified reference protein, such as bovine serum albumin. The BCA assay features compatibility with surfactants and less protein-to-protein variation while the Bradford assay features easy preparation and compatibility with chaotropic and reducing agents. Therefore, the BCA assay is the most popular choice in lysate protein quantitation because of the surfactants in lysate samples.

1.2.2.3 Size-exclusion chromatography

Frequently, it is necessary to remove certain small molecule interference, such as assay incompatible surfactants, buffering agents/salts, excessive reagents, reducing agents, and chaotropic agents, from the lysate sample matrix while retaining its protein content. The size-exclusion chromatography (SEC) is the most efficient technique to perform a buffer exchange, which replaces the entire sample matrix of small molecules with one compatible to downstream processes. The buffer exchange SEC is usually performed in spin column format. The lysate sample is gradually loaded to the resins packed in a spin column, where small molecules will enter the resin matrix and become retained while large molecules will bypass the resin matrix and elute early. Alternative techniques, including molecular filtration using micro-spin filters and dialysis using semi-permeable membrane inside a cassette, are also available. However, compared to SEC, these techniques are usually less effective and more time-consuming.

1.2.2.4 Gel electrophoresis

Within the polyacrylamide gel, the porous gel matrix retains proteins and isolates them from the sample matrix. Driven by electrophoretic forces in the electric field, proteins undergo differential migration and thus separation in the polyacrylamide gel matrix with an appropriate buffer composition according to their size (sodium dodecyl sulfate-polyacrylamide gel electrophoresis, SDS-PAGE), charge (isoelectric focusing-polyacrylamide gel electrophoresis, IEF-PAGE), or a mixed factor of the size, charge, and shape (native polyacrylamide gel electrophoresis, native PAGE). The two-dimensional gel electrophoresis is essentially a serial implementation of IEF-PAGE and SDS-PAGE as orthogonal separation techniques to resolve different proteins in a lysate sample. Gel electrophoresis techniques are used for both analytical and preparative purposes in proteomics research. Analytical protein gel electrophoresis provides an expedited way to visualize the global protein composition of the sample upon colorimetric staining and validate target proteins upon affinity/reaction-based fluorescence detection, which can be used as both an upstream sample preparation checkpoint and supplementary data to the mass spectrometry-based proteomics profiling result. On the other hand, preparative protein gel electrophoresis provides an interactive approach to purify protein analytes and isolate proteins of interest with on-spot verification. The (fluorescence imaging) verified target protein gel bands or spots can be excised and undergo the in-gel protein digestion process to prepare enriched MS-compatible analytes.

1.2.2.5 Affinity capture

When proteins of interest are distinguishable from the rest of proteins within the sample matrix, affinity capture (also known as affinity pull-down or affinity enrichment) techniques can be used to concentrate analytes and simplify the sample matrix. Engineered avidin immobilized

on agarose or magnetic beads can enrich proteins chemically labeled with the biotin affinity tag. Monoclonal antibodies immobilized on magnetic beads can also capture specific target proteins with or without engineered amino acid sequence tags (such as polyhistidine tags) via immunoprecipitation (IP) or protein complex immunoprecipitation (Co-IP) processes.

The biotin-(strept)avidin system serves as the technological core of an enormous number of approaches to analyze and manipulate molecular interactions within complex biological matrices. It has been decades since the discovery and first laboratory application of this tightest protein-ligand binder ($K_d \sim 10^{-15}$ M) that nature ever creates. Yet, its prevalence in biological sciences never seems to diminish.[32,33] Despite nuances among the strategies of its specific applications, the four-step general principle remains the same: (a) Tag the target with biotin, directly or indirectly, covalently or non-covalently. (b) Conjugate the (strept)avidin to a reporter group (for detection) or immobilizing matrix (for isolation), directly or indirectly, covalently or non-covalently. (c) Remove irrelevant species by fractionating a reaction mixture of the tagged target and its complementary (strept)avidin conjugate. (d) Analyze and recover relevant remnants by monitoring and alternating its physicochemical environment, such as photochemical response, pH, redox, etc.

1.2.2.6 Protein digestion

Bottom-up proteomics focuses on the analysis of proteolytic products of protein mixtures. It requires at least one protease to cleave protein analytes into shorter peptides. LC-MS/MS analyzes the resulting complex peptide mixture. Among all naturally occurring or engineered proteases, trypsin is usually the protease of choice for protein digestion in bottom-up proteomics. As a serine protease found in animal small intestines, trypsin is highly specific and only cleaves protein amide bonds at the carboxyl side of arginine and lysine residues except those before the

proline. The resulting tryptic peptides carry either a C-terminal arginine or lysine, both of which are highly basic and easily carry positive charges, leveraging their ionization efficiency and thus detectability by mass spectrometers. Less frequently, alternative proteases such as chymotrypsin, LysC, LysN, AspN, GluC, ArgC, etc. are also used in proteomics.[34]

1.2.2.7 Offline peptide fractionation

In proteomics profiling, a complex lysate digest can be prepared into several fractions before the LC-MS/MS analysis to improve the overall analytical performance. Similar to two-dimensional online LC, offline fractionation also utilizes an additional chromatographic mechanism that is orthogonal or semi-orthogonal to the online reverse-phase chromatography. Popular peptide fractionation methods include strong cation exchange (SCX)[35] chromatography, hydrophilic interaction chromatography (HILIC)[36], and high-pH reversed-phase fractionation (HpH)[37]. These methods have been successfully implemented in large-scale proteomics studies to greatly improve the protein sequence coverage and PTM hits.

1.2.3 Liquid chromatography

In theory, mass spectrometry is perceived as the central technology that allows identification and quantitation of thousands of peptides in a complex proteome digest. Liquid chromatography (LC) is recognized as the indispensable tool that provides high-performance real-time separation of those peptides. HPLC relies on fluidic pumps to pressurize liquid solvents (mobile phase) that carry a sample mixture of analytes through a column packed with a solid sorbent material (stationary phase). Governed by its physicochemical properties, each analyte in the sample interacts differently with the stationary phase as the sample mixture travel through the column. The separation of analytes occurs when the retention differences of analytes within the sorbent matrix are distinguishable. Apart from the column length, the LC separation efficiency

depends on several physical parameters. With a given set of these parameters, the separation efficiency per unit column length can be described as the reciprocal of theoretical plate height which is modeled by the van Deemter equation[38]. For a specific LC column, the smaller the theoretical plate height, the larger the theoretical plate number, and thus the higher the separation efficiency. As exemplified in **Figure 1.4**, the van Deemter equation is a hyperbolic function that predicts the plate height from a linear velocity of mobile phase as the variable and three contributing terms as constants. This model implies it is possible to maximize the separation efficiency for a specific column by just tuning the flow rate of the LC for the optimal linear velocity of the mobile phase.

In practice, LC serves as the last but best step of separation that reduces proteomics sample complexity prior to mass spectrometric analysis. Although the LC separation plays a critical role in bottom-up proteomics for protein/peptide identification and quantitation, LC methods are developed and implemented in a relatively conservative and standardized manner. A typical bottom-up proteomics LC setup consists of a nanoflow LC system, a nanoflow reverse-phase column (usually packed with C18 stationary phase) coupled to a compatible emitter interfacing the ESI inlet, and a binary mobile phase system (usually comprised of 0.1 to 0.2% of formic acid in water as the weak solvent and 0.1 to 0.2% of formic acid in acetonitrile as the strong solvent). Such nanoflow reverse phase LC setups are empirically known, heavily tested, and experimentally justified for the optimal peptide resolution and sensitivity as well as practicability. In addition to the simple reverse-phase LC setup, the multidimensional protein identification technology (MudPIT) was designed to integrate an SCX column and a reversed-phase column in a biphasic LC system for a higher number of resolvable peptides by the mass spectrometer.[39] However, these

multidimensional LC setups gradually become less desired as mass spectrometers become more powerful in separating and resolving ions in the gas phase.

1.2.4 Mass spectrometry instrumentation and implementation in proteomics

The general principle of mass spectrometry (MS) is to generate ions of analytes from samples and measure them via an optimal analytical method. Capable of ionizing analytes, separating, and detecting ions in the gas phase, all mass spectrometers share the basic setup consisting of an ion source where solution-phase molecules are converted to gas-phase ions, mass analyzer(s) where gas-phase ions are separated, and detector(s) where signals from the ions are measured. The readouts of mass spectrometers are *mass-to-charge ratio* (*m/z*) values and their abundance-correlated signal intensities over the period of measurement. MS datasets can provide two types of extracted data: mass spectra and ion chromatograms. While a plot of intensity vs. *m/z* values at a specific time provides a mass spectrum, the plot of ion intensity (sum of intensity/current within a given *m/z* range) vs. time presents an ion chromatogram. Subsequently, mass spectra are generated for compound identification, but chromatograms are derived for analyte quantitation. In addition to the basic MS setup, advanced mass spectrometers are also equipped with collision cell(s), where deliberate gas-phase fragmentation of target ions can take place. These instruments offer multi-stage experimental setups that enable breaking down analyte ions into fragments in the gas phase for further analysis. These setups are defined as tandem mass spectrometry, MS/MS (MS^2), or MS^n, where "n" is the number of fragmentation stages. In bottom-up proteomics, it is the MS/MS that serves as the technical fundamental to sequence peptides and identify proteins.

1.2.4.1 Ion sources

Mass spectrometry is the study of gas-phase ions and their chemistry. The ion source generates these ions. Since the downstream components in a mass spectrometer such as mass analyzers and detectors can only handle charged species, ionic forms of analytes in the gas phase must be created from analyte molecules in solutions. Although a wide range of ionization methods are available, proteomics primarily uses electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) due to their preferred "softness" for ionizing peptide and protein analytes while causing minimal in-source fragmentation of these analyte ions.

ESI is known as the softest ionization technique for solution-to-gas phase conversion of analytes. Due to its ionization softness and universal compatibility to LC systems, ESI is the method of choice in analyzing peptides and proteins.[40] At the ESI source, the pressurized sample solution (either from LC or from a syringe pump) is forced to pass through an electrospray emitter (a needle made of either metal or metal-coated glass) at a flow rate ranging from hundreds of nanoliters per minute to microliters per minute. The emitter is constantly charged and holds an electric potential of 2 to 6 kV relative to the ion inlet. During the ESI process (**Figure 1.5**), the electrically sprayed aerosol, comprised of microscopic droplets, is generated. These droplets, which carry electric charge on the surface, gradually shrink and disintegrate due to the evaporation of solvent and desorption of ions from the solution. Analyte ions are generated via two mechanisms: Coulombic fission and surface evaporation. The former mechanism assumes that the increase of charge density of a droplet, due to solvent evaporation, causes the droplet to split into multiple smaller ones, which ultimately form individual ions in the gas phase. The latter assumes that the increase of charge density of a droplet, resulting from solvent evaporation, escalates Coulombic repulsion, which eventually overcomes the surface tension of the liquid, causing the

release of ions from the surface of droplets. This process is usually accelerated by heat and countercurrent streams of nitrogen gas at microliter flow rates of sample infusion.

MALDI ionizes the dried samples mixed with a suitable crystalline matrix such as sinapinic acid and ferulic acid on a metal plate via pulsed laser (**Figure 1.5**). It is known only to generate singly charged ions. In proteomics, MALDI-MS is normally used for the fast identification of reconstituted protein samples isolated from gel electrophoresis, size exclusion chromatography, and ion exchange chromatography.[41] Unlike ESI, which interfaces with LC systems at the atmospheric pressure, the MALDI source usually requires the prepared sample matrix plate placed in a vacuum chamber where the ionization takes place. The potential uneven distribution of analytes on the matrix plate makes MALDI a less quantitative ionization method compared to ESI. Therefore, the application of MALDI is normally limited to the qualitative protein analysis in the top-down proteomics while the ESI is more versatile.

1.2.4.2 Mass analyzers

In a mass spectrometer, the mass analyzer separates ions in the gas phase before the detector measures the *m/z* values of these ions. The separation of gas-phase ions can be either spatial or temporal, according to the physics of the mass analyzer. The resolving power of a mass analyzer describes its ability to separate ions with similar *m/z* values. The higher the resolving power, the smaller the distinguishable differences among *m/z* values. Quadrupole (Q) and ion trap are two types of low-resolving power mass analyzers used in proteomics. Time-Of-Flight (TOF), Fourier-transform ion cyclotron resonance (FT-ICR), and Orbitrap are three types of high-resolving power mass analyzers used in proteomics. Both quadrupole and TOF separate ions in space. The ion separation occurs in a quadrupole via its alteration of the electric field that enforces selective transmission of ions within a defined *m/z* window. Therefore, a quadrupole mass analyzer

is also referred to as a quadrupole mass filter or mass selector. The ion separation occurs in a TOF mass analyzer via the amplification of the *m/z* dependent ion differential flight distances in the vacuum after the initial acceleration in the electric field. The TOF mass analyzer features the intrinsic compatibility to the MALDI source and a wide measurable *m/z* range at a high spectrum acquisition rate; its use is mostly preferred in top-down proteomics. By trapping ions in an alternating electric field in the vacuum, the ion trap, FT-ICR, and Orbitrap separate ions in time. Detection of trapped ions in an ion trap requires the selective release of ions with specific *m/z* values and the physical contact of these ions with the detector. However, the detection of trapped ions in the FT-ICR and Orbitrap does not require the physical contact of any ion with the detector; its detection is achieved via electromagnetic induction. In practice, the FT-ICR is known for its highest resolving power but low scan rate, which is only suitable for top-down proteomics analysis. Orbitrap is becoming the gold standard for proteomics due to its tunable performance balance between the resolving power and scan rate, which can be used in both top-down and bottom-up proteomics.

1.2.4.3 Gas-phase peptide fragmentation and tandem mass spectrometry

The MS/MS process can be simply perceived as two-stage mass analysis of analyte ions. This process thus involves two mass analyses, with the first mass analysis selecting ions within a specific *m/z* window. This *m/z* value is characteristic of a pre-defined analyte precursor from the sample mixture. These precursor ions then pass through a gas-pressurized compartment where they are activated and break into fragments that are also known as product ions. The fragmentation can be achieved via different mechanisms. Collision-induced dissociation (CID; also known as collisional activation dissociation, CAD) and higher-energy collisional dissociation (HCD) are two most used mechanisms, where accelerated precursor ions collide with neutral gas molecules to

accumulate thermodynamic energy for their dissociation. The resulting product ions are further subjected to the second mass analysis, following by ion detection for compiling a structure-specific mass spectrum for the precursor.

Unlike those of proteins and small metabolites, fragmentation patterns of peptides that provide the sequence information are well documented and highly predictable. The trait makes the bottom-up proteomics results more credible and reproducible than others. The interpretation of a peptide MS/MS spectrum involves assignments of six types of fragment ions generated from the peptide backbone. When the cleavage of an amide bond on a positively charged peptide occurs, b and y ions are produced. The apparent cleavage of a Cα-C bond annotates a and x ions. When the cleavage of an N-Cα bond occurs, c and z ions are produced. The a, b, and c ions have a positive charge on the N-terminus while the x, y, and z ions have a positive charge on the C-terminus. The number of amino acids contained in the fragment ion is noted in a subscript, for instance, b_3 as the N-terminal ion produced by the cleavage of the 3rd amide bond counting from the N-terminus, and y_6 as the C-terminal ion produced by the cleavage of the 6th amide counting from the C-terminus (**Scheme 1.1**). Systematic analysis of mass spectra containing information on these product ions forms the technological basics of mass spectrometry-based peptide sequencing.

In addition to CID and HCD, other fragmentation mechanisms such as electron-transfer dissociation (ETD), electron capture dissociation (ECD), and ultraviolet photodissociation (UVPD) are also used in proteomics for studies of PTMs, intact proteins, and cross-linked protein complexes. ETD and ECD fragmentation mechanisms are known to preserve PTM information better than CID and HCD. In comparison to CID and HCD that primarily produce b and y ions from peptides, ETD and ECD can produce a predictable, homologous series of c and z ions that are absent in CID or HCD-based MS/MS spectra.

1.2.4.4 Modes of spectral data acquisition in bottom-up proteomics

The bottom-up proteomics pipeline begins with proteomics profiling experiments to characterize complex proteome samples and discover proteins of interest from identification and quantitation of individual protein components. The data-dependent acquisition (DDA, **Figure 1.6**) mode is the classic LC-MS/MS method for measuring protein samples in a proteomic profiling experiment.[42,43] In the LC-MS/MS profiling experiment, LC column-separated tryptic peptides are introduced to the mass spectrometer and converted to protonated gas-phase ions at the ESI source. The mass spectrometer loops the measurement of these peptide ions in duty cycles. In each duty cycle, the instrument first performs a quick MS1 survey scan for all the detectable ions. Second, the instrument identifies "top N" (according to the pre-set "N" value in the instrumental method) precursor ions with the highest intensities as targets for gas-phase fragmentation and MS2 analyses. Third, the instrument performs a series of MS2 scans for all these target ions. During each MS2 scan, the instrument isolates, fragments, and measures the product ions of each pre-selected target ion, generating target-specific MS2 spectra. Moreover, programmable DDA filters (such as charge state inclusion/exclusion, dynamic exclusion, and isotope pattern matching) are used routinely to maximize the detectability and MS2 analyses of peptide analytes. However, due to the stochastic ion sampling, limited scan rates of mass spectrometers and inadequate peak capacities of LC[44], DDA data usually fails to provide a complete picture of the sample.

In recent years, data-independent acquisition (DIA) has been gaining popularity[45-49], thanks to faster computers and sophisticated data mining algorithms[50-54]. In contrast to DDA, DIA (**Figure 1.6**) does not rely on the MS1 survey scans that trigger the selection of detected precursor ions for gas-phase fragmentation and subsequent MS2 scans of their product ions. In a DIA mode, the mass spectrometer is programmed to alternate periodically either the collision energy (all-ion-

fragmentation, AIF[55]; MSE[56]) or precursor ion isolation window (MSX-DIA[57,58], SWATH[59]) so that each duty cycle can cover a broad range of spectral information. However, such information is usually not immediately available and requires in-depth post-acquisition processing of resulting DIA datasets containing complex spectra of fragment ions.[50,60] While the analysis of DIA data can be tricky, each dataset provides a more complete and less biased depiction of the sample.[61] When incorporated with internal standard peptides for retention time normalization, a DIA method was shown to outperform a typical DDA method in both the number of consistently identified peptides across multiple measurements and quantitation of proteins with various abundance.[62] Furthermore, archived DIA datasets hold more value than DDA datasets in retrospective analyses[60,63] when the knowledge of previously-unknown analytes becomes available. However, DIA methods lack the ability of generating high-quality precursor-specific MS2 spectra, thus may suffer from poorer specificity of identified peptides than that of DDA methods. Because every scan of DIA contains multiple groups of fragment ions (corresponding to different peptidyl precursor ions) recorded on a single MS2 spectrum in a convoluted manner while DDA scans are designed only to measure the fragment ions generated from a single defined precursor ion in principle.

In the downstream segment of the bottom-up proteomics pipeline, protein targets of interest are usually analyzed as tryptic peptides in a target-specific manner that ignores or removes everything other than the target analytes from the sample. The targeted proteomics relies on two acquisition modes known as selected reaction monitoring (SRM; also known as multiple reaction monitoring, MRM) on the triple quadrupole (QqQ) tandem mass spectrometer and parallel reaction monitoring (PRM) on the quadrupole-orbitrap (Q-Orbi) tandem mass spectrometer. When a QqQ instrument operates in the SRM mode (**Figure 1.7**), its first mass analyzer (Q1) selectively transmits precursor ions of a particular *m/z* corresponding to a pre-selected target analyte. These

precursor ions are then subject to CID in a collision cell (q2) filled with neutral gas regulated with proper instrumental settings. This process produces a variety of fragment ions of the target analyte. However, only a few pre-selected fragment ions with high intensity and specificity are, once again, selectively transmitted through the second mass analyzer (Q3). The transmitted ions finally reach the detector in the mass spectrometer, and detected signals are recorded as an ion chromatogram for a precursor-fragment ion pair, which is known as a transition in the SRM MS experiment. Transitions are pre-determined during the experiment design stage. Notably, new QqQ mass spectrometers are faster and can analyze a larger number of transitions within a duty cycle. The duration of a duty cycle is defined as the cycle time, while the time the instrument spends on scanning each transition is defined as the dwell time. The cycle time heavily depends on the LC gradient profile, which affects the chromatogram peak widths of peptides because a good quantitation method usually implements 12 to 20 data points per LC peak and 3 to 5 transitions per peptide.[64] Moreover, with robust LC instrumentation and known LC retention profiles of target peptides, the SRM acquisition can be scheduled for a limited number of expected transitions at each specific time window. This scheduled SRM method can greatly leverage the total number of transitions per run. Overall, selective ion transmission features SRM MS analysis with excellent limits of quantitation (LOQs), high specificity, and a wide dynamic range at a lower cost.

In contrast to SRM, the PRM mode is the next-generation ion monitoring technique exclusive to high-resolving power and high-mass accuracy hybrid tandem mass spectrometers like quadrupole-orbitrap.[65-68] The principle of PRM MS is comparable to SRM MS (**Figure 1.7**). Briefly, the quadrupole mass analyzer of these tandem instruments selects precursor ions. These precursor ions then undergo high-energy collisional dissociation (HCD) in a collision cell**.** Compared to CID, HCD generates a broader spectrum of fragment ions, which provides a higher

global detectability of peptide analytes.[69] These fragment ions are transmitted to an orbitrap mass analyzer for gas-phase separation and detection at MS2 level with high resolving power and high mass accuracy. Unlike the SRM, which records the signal for one fragment ion of the selected precursor at a time, the PRM simultaneously detects a full range of fragment ions of one precursor. Therefore, the number of fragment ions does not restrict the speed of such a full scan, which is instead limited by the orbitrap mass analyzer's resolving power. The LOQ of analytes in the PRM mode can be improved by summing intensities of multiple fragment ions, which makes up the sensitivity loss due to compromised transmission of individual fragment ions (compared to SRM). Moreover, the full scan of fragment ions can eliminate the need for selecting best precursor-to-fragment ion pairs before LC-MS/MS measurements. For the PRM data processing, the mass accuracy (in ppm) and isotope distribution patterns of each measured fragment ion can be incorporated as part of data-refinement algorithms to minimize the background interference and false detection. The implementation of PRM-based quantitation methods can be easier than SRM-based methods due to the fact that targeted proteomics methods are usually built based proteomics profiling data, which are mostly acquired on the same Orbitrap-based mass spectrometers used in PRM MS experiments due to the growing trend of Orbitrap-based proteomics practice. Overall, in addition to its SRM-comparable target quantification capability, the PRM technique offers more analyte multiplexity and an easier LC-MS/MS-based assay development workflow.

1.2.5 Bioinformatics of database search and peptide identification

The rapid advancement of mass spectrometry-based proteomics has shifted large-scale studies of protein sequences from the genomics/transcriptomics-based prediction towards direct analysis of complex protein mixtures. The practice of mass spectrometry-based protein identification also relies heavily on the development of computer algorithms and software tools.

These software tools identify proteins by referring to and searching sequence databases acquired from the genomics sequence infrastructure. In a typical bottom-up proteomics profiling experiment, MS/MS spectra are acquired from peptides generated from trypsin digestion of complex protein mixtures. These MS/MS spectra are stored as two-dimensional arrays of signal intensity (integers) versus *m/z* (floating point numbers) values. Each of these MS/MS spectral datasets is also tagged with its corresponding precursor ion *m/z*, charge state, and retention time on the LC column. In general, a computer program splits the database search process into four steps to assign a peptide identity to such a MS/MS spectrum. These four steps are input pre-processing, *in silico* protein digestion, peptide mass filtering, and peptide-spectrum matching/scoring. (**Figure 1.8**)

1.2.5.1 Input pre-processing

A database search program initiates a new task by reading and extracting information from two pieces of input data: the protein sequence database predicted from genomics and raw MS/MS data (usually acquired by an HR/AM mass spectrometer in DDA mode). The former is stored in FASTA format as a single text-based file compiling all known protein sequences from the proteome of a specified organism. For each protein entry, the first line in a FASTA file starts with a”>” (greater-than) character that followed by the entry title consisting of the protein name and comments parsed by other special characters. The second line is the actual protein sequence itself in the standard single letter code for amino acids. The raw LC-MS/MS data are stored in proprietary formats (e.g., RAW, and WIFF) that are convertible to open-source formats (e.g., mzML, and MGF). The program either directly reads raw data files in their proprietary formats or requires third party software for the conversion to an open-source format. After the import of raw data, the database search program separates MS/MS spectra from MS survey spectra and performs

spectral data reduction with a series of calculations that convert the raw MS/MS spectra to search-compatible ones. These calculations include centroiding, mass rounding, de-isotoping, and charge state reduction. The centroiding refers to calculations that extract apex *m/z* values of continuously sampled data points in the interval recording (profile or continuum) mode of the mass spectrometer. The mass rounding refers to calculations that round the *m/z* values to less precise figures according to a pre-defined mass accuracy threshold. The de-isotoping refers to the practice that merges multiple *m/z* peaks and isotope clusters that feature the same fragment ion in different charge states on the sample MS/MS spectrum. The charge state reduction is always performed as part of the de-isotoping, which preserves only the *m/z* peaks of singly charged fragment ions on processed MS/MS spectra. These pre-search MS/MS spectrum processing measures can greatly facilitate down-stream search algorithms by reducing data complexity and increase the overall computational efficiency.

1.2.5.2 *in silico* protein digestion

Similar to the process of the actual enzymatic digestion of proteins, the *in silico* protein digestion refers to calculating and indexing theoretical *m/z* values from protein sequences fetched from the database. For each protein, the digestion algorithm predicts a list of peptides according to the specified protease and its empirical enzymatic cleavage sites (such as cleavage at the C-terminal side of lysine or arginine for trypsin) as well as a pre-defined number of missed cleavage (usually 1 to 3). Theoretical masses of these peptides are then calculated from their predictable elemental compositions based on their amino acid residues, pre-defined chemical modifications, PTMs, and secondary fragment ions with neutral losses. Additional constraints such as maximum and minimum peptide lengths/masses are also pre-defined as included in configurable search

parameters. The resulting theoretical *m/z* values are compiled as theoretical MS/MS spectra for later uses.

1.2.5.3 Peptide mass filtering

Before the search program can match an experimental MS/MS spectrum to a theoretical one and score it for a peptide hit, the theoretical spectra are filtered at the MS level. For each experimental MS/MS spectrum tagged with its precursor mass, the filtering algorithm extracts theoretical MS/MS spectra that have identical masses within a pre-defined mass accuracy threshold (usually ± 5 to 10 ppm). The search program may perform additional search engine-specific filtering that dynamically removes low-intensity peaks from an experimental MS/MS spectrum to improve its overall spectrum quality.

1.2.5.4 Peptide-spectrum matching and scoring

The peptide-spectrum matching (PSM) algorithm is the core algorithm that compares candidate sequences to experimental MS/MS spectra with appropriate precursor masses. Different spectra may have different maximum numbers of detectable fragment ion masses, numbers of peaks with different quality, and signal-to-noise ratios. Therefore, successful identification of an observed MS/MS spectrum depends on several factors, including the number of matching peaks, mass accuracy, overall spectrum quality, and the uniqueness of the underlying peptide within the whole proteome. To report meaningful PSM results, search engines usually adopt mathematical and statistical models to score peptide-spectrum matches (PSMs) and systematically describe the statistical significance and confidence of these PSMs before they report them in the final result. A target-decoy strategy,[70] where a decoy database (consisting of either reverse or scrambled protein sequences) is generated from the input database and used together for PSM, is implemented to estimate the false-discovery rate (FDR, usually 1% as the cutoff) for the quality control and last-

step filtering of the final result. Different search engines implement the PSM scoring system differently. **Table 1.1** summarizes some key features of five popular search engines that are heavily used in the field.

1.3 Tools and Strategies in Chemical Proteomics

The mass spectrometric analysis in most proteomics studies has extended beyond the profiling of protein expression. The investigation and utilization of modified proteins and peptides have become major topics in proteomics research. A modification can be broadly defined as an observable positive delta mass (mass increase) on the protein or peptide. It can be a result of either the chemical alteration or biological manipulation of proteins as different modifications are introduced to or exist intrinsically within a proteome. Some modifications can be products of chemical probes that label a specific number of target proteins. Some can be metabolic labels or chemical tags introduced to proteins and peptides during the sample preparation to achieve analytical objectives. Others may simply exist naturally as part of cell signaling and regulation processes where enzymes and biochemical pathways are activated. Although conducted for diverse chemical or biological principles, these modification-specific proteomics studies share similarities in multiple analytical perspectives of sample preparation, analyte detection, and data interpretation.

Chemical proteomics or chemoproteomics can be defined as a proteomics-based systematic approach to study the interaction between small molecule compounds and proteins in complex biological systems. In contrast to the conventional "one-on-one" strategy used in studying small molecule-protein interactions such as ligand binding assays, chemical proteomics emphasizes multiplexed protein analysis with the inclusion of the biochemical complexity in the experimental model, where the small molecule-protein interactions take place. Therefore, compared to traditional methods in chemical biology and drug discovery studies, chemical proteomics profiling is more suitable for discovering protein targets of small molecule compounds, measuring target engagement, understanding mechanisms of action, and evaluating off-target effects. Chemical

proteomics relies on chemical probes that introduce detectable chemical modification and/or distinguishable analytical traits of proteins that interact with the small molecule of interest. Consequently, the preparation of probe-treated protein samples, detection of probe-modified peptides, and analysis of probe-reacted protein depends on a dedicated set of chemical tools and analytical strategies.

1.3.1 Compound-centric vs. activity-based probes

Chemical proteomics studies usually involve one of two probing strategies (**Figure 1.8**). The compound-centric strategy features the use of chemically modified, either tagged[71] or immobilized,[72] compounds of interest as either covalent or non-covalent chemical probes. These chemical probes are introduced directly as baits to capture target proteins from a complex proteome. In contrast, the activity-based strategy depends on covalent probes. These activity-based probes (ABPs) are capable of irreversibly binding, reacting with, and labeling target proteins from a complex proteome, which is pre-treated with compounds of interests. The subsequent elucidation of target activity relies on the measurement of probe-target adducts in competitive binding assays.

Presumably, the compound-centric strategy is straightforward for studying the proteome-wide activity of a known drug molecule. It mandates a boutique probe created by installing a reporting group on the drug molecule. However, the construction of a drug-derived probe often bears high synthetic costs. The application of such a compound-centric probe also suffers from the consequence of chemical modifications; any structural changes made to the drug molecule may significantly alter the parent molecule's potency and selectivity profile.[73] The more sophisticated strategy is activity-based chemical proteomics, which emphasizes the use of a covalent probe. Activity-based competitive chemical proteomics, also technically known as competitive activity-based protein profiling (ABPP), offers a versatile bioanalytical platform.[74,75] Such a platform can

effectively decipher proteome-wide actions of different underivatized drugs and other reactive molecules like environmental toxins and reactive metabolites from the human microbiota.[76-79] Opposing to the design of compound-centric probes, ABPs are designed to show little or no specificity to proteins. They capture proteome-wide "snapshots" visualizing drug-protein interactions by permanently occupying available active sites post drug treatment on the model proteome. As illustrated in **Figure 1.8**, a competitive ABPP platform depends on its ABP to measure the proteome-wide action of an underivatized drug.[80,81] Competitive ABPP's distinctive technological advantage is that a single ABP with broad proteome coverage can establish a versatile analytical platform capable of evaluating different drugs or drug candidates on different subsets of a single proteome. Therefore, ABPs with a broad spectrum of reactivity can fully unleash the enormous potential of competitive ABPP technology.

1.3.2 Chemical tagging on peptides

The quantitative analysis of *in vitro* drug dose response is challenging. It demands not only identification but also accurate and precise quantitation of biochemically altered proteins from proteome samples treated at various drug concentrations. Fortunately, several types of peptide derivatization reagents are available for label-based quantitation. These versatile derivatization reagents tag multiple protein digest samples that are combined for simultaneous analysis, which enhance both the accuracy and sample throughput. These chemical tags can be either stable isotope-based or non-isotopic.

Stable isotope-based peptide tagging involves either isobaric[82-84] or mass-difference[85-94] derivatization reagents. These reagents tag peptides and incorporate stable isotope labels differentially. In comparison to mass-difference derivatization reagents, isobaric reagents are more popular because they are commercially available, more reliable, and constantly being updated for

additional multiplexity: 4-plex iTRAQ,[83] 6-plex TMT[84], 8-plex iTRAQ[95], 11-plex TMT[96,97], and 16-plex TMT[98]. Using tandem mass tag (TMT™) reagents as an example, peptides derivatized with isobaric reagents have the same precursor ion masses but produce product ions with differentiable masses only after gas-phase fragmentation (**Figure 1.10A**). Therefore, tandem mass tags enable the concurrent measurements of chemical proteomics samples. The LC-MS/MS measurements of pooled TMT-labeled samples minimizes unwanted run-to-run variations in target quantitation. Because these tandem mass tags multiply the abundance of a peptide precursor ion by the number of combined samples. This analyte abundance boost has little effect on the spectrometric complexity during the MS precursor ion scan of a DDA profiling experiment. Only the reporter ions are measured for peptide quantitation during the MS/MS product ion scan, which significantly mitigates chemical interference. Besides, signal intensities of sequence ions from different samples are additive. This feature ensures that combining multiple samples for a single LC-MS/MS run enhances the detectability of peptide sequence ions and has a minimal negative impact on the sample matrix. As illustrated in **Figure 1.10B**, the 10-plex TMT-based chemical proteomics workflow can provide a powerful shortcut to quantify dose responses of small molecule compounds of interest within a complex proteome.

It is also possible to use mass-difference derivatization reagents for incorporating peptides with stable isotope labels.[86,87,90,92,94] For instance, the reductive methylation of peptidyl amines is a commonly practiced method for multiplexed proteomic quantitation.[85,88,89,91,93] It features a simpler derivatization procedure, lower cost,[91] and the potential for automation.[88,99] However, it can only be applied for labeling up to five different samples. The quantitation of methylated peptides is based on fragment ions generated from tagged peptides with different precursor masses., resulting in a relatively higher chemical background and spectrometric complexity.

Non-isotopic derivatization reagents leverage the analytical performance of a chemoproteomics platform in a slightly different way. These reagents are designed to provide a cost-effective solution to improve the sample throughput in targeted proteomics quantitation rather than analyte multiplexity in untargeted proteomics profiling. This technology is proposed as ultra-throughput MRM MS, or uMRM MS.[100,101] This method re-allocates the scan capacity of fast contemporary QqQ instruments from a large number of pre-selected peptides in a single sample[102] to several key peptides of interest from a large number of samples in a single experiment.[100,101] As demonstrated on **Figure 1.11**, once integrated with stable isotope by amino acids in cell culture (SILAC), the open-source uMRM MS technology can also greatly facilitate the development of chemical probes and chemical proteomics by enhancing the quantitation performance in the scaling analysis of proteome-wide reactivity of activity-based chemical probes.[103]

1.3.3 Metabolic labeling on proteins

Metabolic labeling methods incorporate stable isotope labels biologically into a model proteome during the cell culture (SILAC) or small animal feeding process (stable isotope labeling by amino acids in mammals, SILAM).[104] These powerful labeling methods are residue-specific and rely on the natural metabolism of a living system to produce heavy copies of proteins.[105] Although commonly used in a heavy/light duplex fashion, up to five samples can be pooled for concurrent proteomic quantitation.[106] Compared to chemical tagging, the most significant advantage of metabolic labeling is the potential of introducing the labeled proteins at an earlier stage of proteomic sample preparation workflow. Thus, the reduction of sample complexity at the protein level can be implemented for high accuracy and precision proteome quantification, with little differential protein loss.[107] By only supplying heavy amino acid feed to an organism, organism-wide labeling is achievable. The SILAC technology is widely used for investigating

disease mechanisms and identifying biomarkers in human samples.[108-112] Similar to the mass-difference tagging, the use of metabolic labeling suffers from analyte dilution and increased sample complexity. Nevertheless, as demonstrated on **Figure 1.11**, once integrated with the open-source uMRM MS technology, SILAC the can also greatly facilitate the development of chemical probes and chemical proteomics by enhancing the quantitation performance in the scaling analysis of proteome-wide reactivity of activity-based chemical probes.[103]

1.4 Challenges and Opportunities in Chemical Proteomics

1.4.1 Alternative bio-orthogonal affinity tags and trypsin-resistant affinity binders

In most cases, biotin-(strept)avidin system-based techniques are versatile and fit seamlessly in their applicable experimental workflow. However, when being used for affinity enrichment of probe-reacted proteins, this pair of binders often delivers poor enrichment efficiency due to the mandatory harsh conditions for releasing captured targets, such as prolonged heating and use of chaotropic agents at an extremely low pH for analyte elution. While these harsh conditions can disturb the strong biotin-(strept)avidin interaction, they may also inflict undesired damages on both the solid-matrices and captured analytes, thus increasing the sample's complexity and decreasing the analyte's recovery. In comparison to biotin ($K_d \sim 10^{-15}$ M), the desthiobiotin has been developed as a synthetic biotin analog with a deliberately-reduced avidin affinity ($K_d \sim 10^{-11}$ M) to afford less chemical disruption for its release.[113] The desthiobiotin affinity tag also has its unique advantages as a xenobiotic tag, which is chemically unique and bio-orthogonal with a biological system. Its mass signature is distinguishable by mass spectrometric analysis. It also is not part of any biochemical reactions. On the other hand, biotin-specific antibodies with lower affinities ($K_d \sim 10^{-8}$ M) have also been developed to serve as an affinity enrichment platform for higher analyte recovery.[114]

Nevertheless, protein-based macromolecular binders share a significant drawback. That is chemical orthogonality. As proteins, these macromolecular binders are prone to biochemical manipulations that affect proteins. Trypsin digestion is often implemented in technically distinct ways (**Figure 1.12**) in a chemical proteomics profiling workflow.[115] It is practiced as an adaptive

step within the sample processing workflow according to specific analytical strategies, which would provide distinct advantages and drawbacks. Compared to the other two methods, the on-bead digestion approach provides an uncompromising shortcut to protein-level enrichment. Because the captured probe-protein adducts do not leave the solid matrix until the very end, fewer steps and less analyte loss are expected. However, in practice, there is always an intangible trade-off between the trypsin digestion efficiency and contamination from tryptic (strept)avidin peptides. Even after meticulous optimization, an on-bead digestion protocol frequently leads to an overwhelming number of either miscleaved analyte peptides or (strept)avidin contaminant peptides. In some applications, co-existence of the enrichment media and trypsin is required, particularly known as the on-surface (on-bead) trypsin digestion strategy[115,116]. In this case, trypsin's proteolytic activity can also affect the macromolecular binders, which causes inevitable contamination and loss of analytes.

In recent years, the significance of ligand-binding capability of oligonucleotides has attracted a tremendous amount of attention. Due to several substantial advantages over antibodies, aptamers have gradually gained ground in clinical applications.[117] In comparison to antibodies that are biologically-produced large-sized (150kDa for IgG made in animals or cell lines) protein molecules, aptamers are chemically-synthesized small-sized oligonucleotides (24kDa for 80-NTs ssDNA, made on automated chemical synthesizer) that are easy to create, use, and manipulate. Therefore, it is possible and will be beneficial to develop alternative affinity-capturing platforms consisting of immobilized aptamers and non-biotin affinity tag-based chemical probes. These novel affinity capturing systems may provide superior flexibility, robustness, compatibility, and efficiency for affinity pull-down of target proteins labeled by chemical probes.

1.4.2 Reactivity and selectivity of chemical probes

In contrast to ideal covalent drugs that react with target proteins selectively and sensitively, ideal chemical probes may occupy both ends of the selectivity spectrum. While a successful compound-centric probe must recreate the high selectivity of its drug template, a versatile ABP would react with a wide range of proteins at their active sites, which are targeted respectively by various drugs to be analyzed. Covalent inhibitors and drugs have prolonged selective engagement with their targets and improved pharmacodynamic properties, but they react unavoidably with off-target proteins, leading to toxicity and safety concerns.[118] During the drug development, chemical proteomics are increasingly used to reveal selective covalent inhibitors early, identify toxicity liabilities, and help mitigate the risk of late-stage failures.[74,119]

ABPs are also uniquely useful in chemical proteomics for measuring the proteome-wide action of all sorts of reactive molecules other than covalent drugs to understand their biological consequences.[74,79,119,120] These molecules include environmental toxins, natural products, reactive metabolites from the human microbiota. At the low end of the selectivity spectrum, broadly reactive ABPs[77,78,121] are used as measurement probes in ABPP platforms that can be flexibly adapted for the (re)activity analysis of underivatized reactive molecules. The platform adaptability relies on the reactivity coverage of the measurement probe—the number of proteins and the type of amino acid residues with which the probe reacts. The broader the coverage, the higher the adaptability.

It is challenging for a warhead to cover a large fraction of a proteome reactivity's domain. Individual proteins differ in molecular composition and structure. Their reactions with small molecules are distinct from site to site, domain to domain, and protein to protein. The diversity of these reactions further increases due to the large number and abundance range of proteins in a

proteome. When designed to be broadly reactive, an ABP does not carry a binding/recognizing group and depends on its warhead for its reactivity and proteome coverage.[122-124] Exemplified by iodoacetamide (IA)-alkyne, the IA warhead reactivity represents the proteome coverage of a class of broadly reactive probes,[76] which react with large numbers of proteins in a proteome, but primarily targeting cysteine residues.[77,119] On the other hand, warheads that target multiple amino acid residues are also available. For instance, ABPs that are developed on the acrylate warhead[125] or sulfonyl fluoride[126] warhead are utilized to target nucleophiles, including amino, hydroxyl, and thiol groups. Although being useful in some cases, probes based on these warheads may limit their applications due to storage, sample compatibility, and experimental reproducibility issues caused by their cytotoxicity, background reactions, and instability. The large extent of nonspecific background reactions usually compromises competition assay-based proteomics analysis of the probe-reacted proteins. Therefore, novel warheads with moderate electrophilicity and reaction rates but a high coverage of residue types are still in high demand for developing activity-based measurement probes. These probes are expected to lessen background reactions and establish more versatile platforms of competitive (re)activity-based proteomics.

1.4.3 Bioinformatics for modification-specific proteomics

Chemical proteomics utilizes chemical probes to study biological functions of proteins and small molecule-protein interactions within a highly complex biological sample. Direct LC-MS/MS analysis of a total protein digest prior to any offline analyte separation is typically futile because of the competition between analytes and interfering non-target peptides for ionization and stochastic sampling resource in the mass spectrometer. Fortunately, only a fraction of peptides are actual analytes that should be selected for LC-MS/MS analysis. The selection/enrichment of these analytes is achieved via the affinity tagging approach, where only peptides modified by the chemical probe that carry an affinity tag are retained for further analysis. This intrinsic trait has shifted the analytical focus from all peptides onto probe-modified peptides, which characterizes chemical proteomics as modification-specific proteomics akin to large-scale PTM studies.

It is challenging to analyze modification-specific proteomics data. Conventional proteomics database search engines handle search requests for modified peptides by expanding the peptide sequence pool using a list of modification masses. These modification masses are set as search parameters before the search workflow initiates. When the numbers and locations of specific chemical modification are unknown (known as a variable modification), this approach has to enumerate all possible instances of such modification on native peptides. It also appends modified peptides to the native ones to create an exponentially larger peptide sequence pool as the search space. Problems occur when multiple variable modifications are introduced because the inclusion of combinatory modified peptide variants artificially inflates the search space with more isobaric peptide masses, resulting in a much larger candidate peptide list. Compared to search without modifications, the subsequent iterative peptide-spectrum matching and scoring algorithm using such an inflated candidate peptide list usually results in increased computational cost,

elevated error rates, and compromised sensitivity. Moreover, some chemically modified peptides may undergo secondary chemical reactions during sample preparation or LC-MS/MS analysis. Others may simply be unknown or mispredicted during exploration of novel chemical probes and biochemical reactions. It is also possible that modified peptides will be less detectable due to their erratic gas-phase behaviors such as low ionization efficiency, poor fragmentation, and unusual fragmentation patterns. On the other hand, existing software packages are developed ubiquitously for analysis of all peptides and proteins. They often offer limited support for post-search processing of modification-specific data. These packages usually frustrate end-users by presenting proteomics profiling reports adulterated with a large volume of information on native proteins and peptides that are irrelevant to the research objectives. Therefore, novel search algorithms and post-search data processing tools remain to be developed. Together with novel chemical probes and analytical innovations, these bioinformatics tools will greatly accelerate advancement of chemical proteomics field.

1.5 Chapter 1 Figures

Figure 1.1 The omics pyramid.

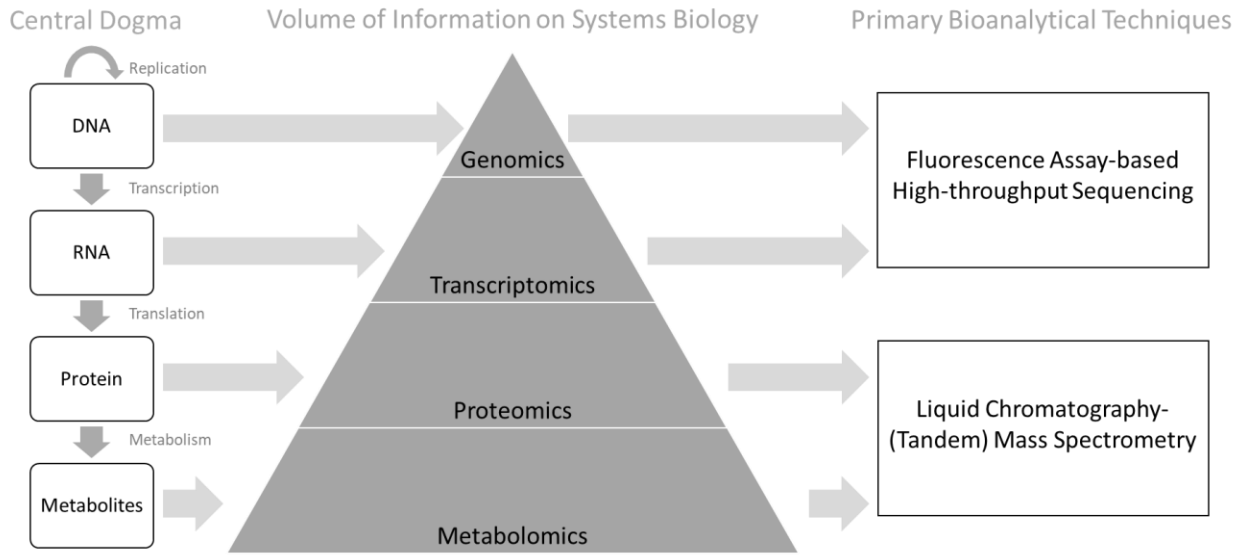
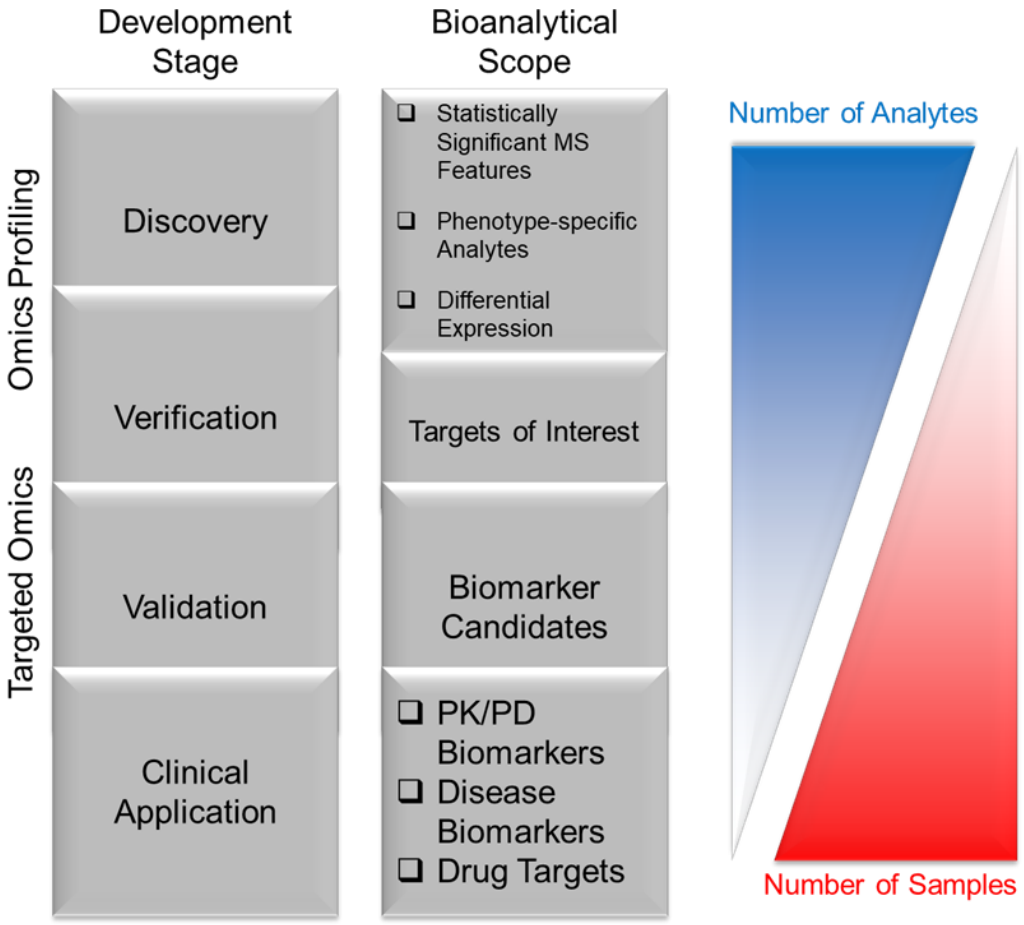


Figure 1.2 Overview of omics pipeline.



Note: Figure adapted with permission from (Wang, L., McShane, A. J., Castillo, M. J. & Yao, X. in Proteomic and Metabolomic Approaches to Biomarker Discovery (Second Edition) (eds Haleem J. Issaq & Timothy D. Veenstra) 261-288 (Academic Press, 2020).). Copyright (2020) Elsevier Inc.

Figure 1.3 Overview of the bottom-up proteomics workflow

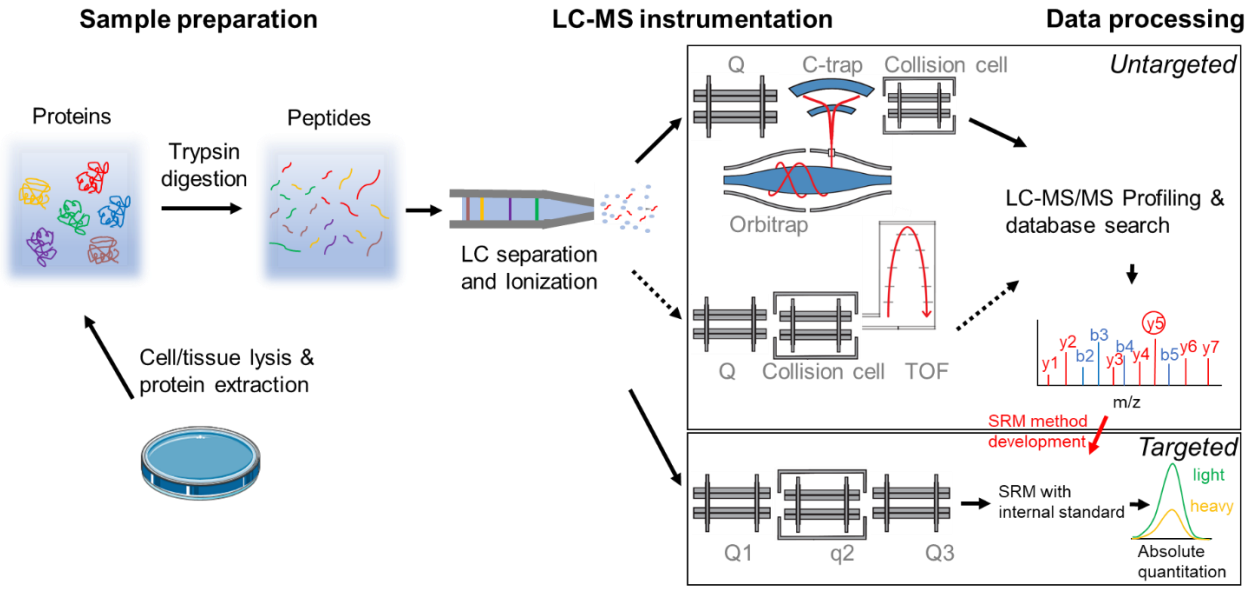


Figure 1.4 Example van Deemter plot with the equation and contributing terms.

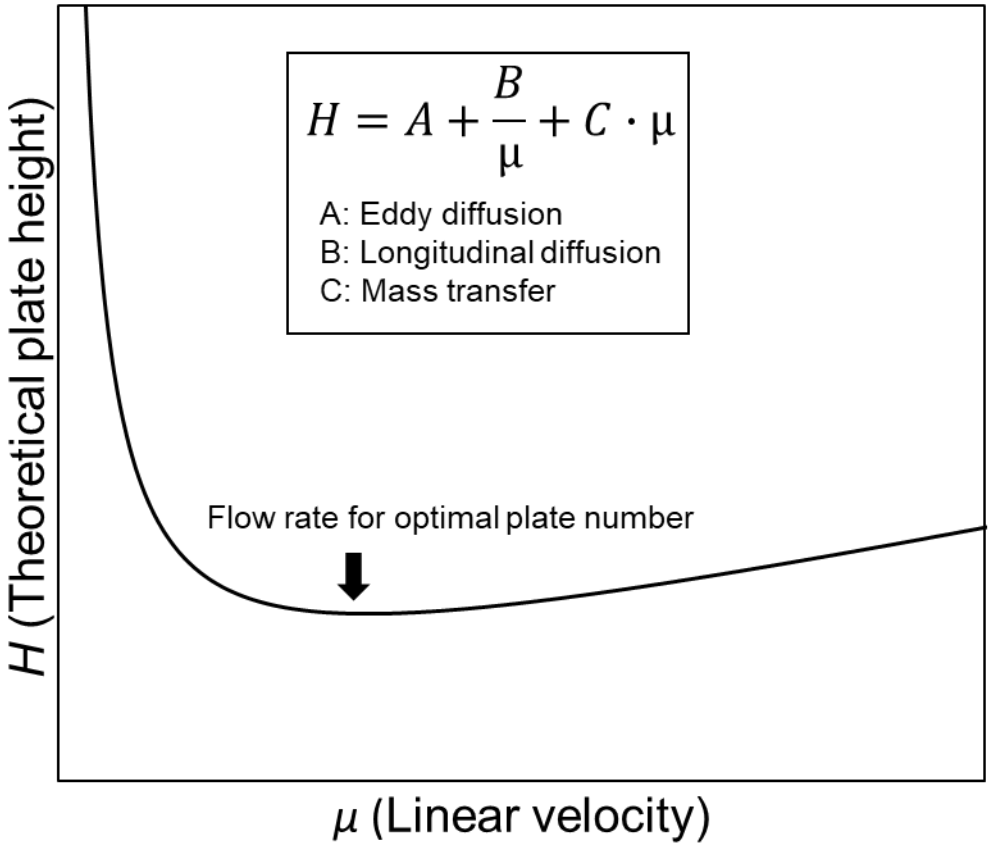


Figure 1.5 Schematic overview of ESI and MALDI sources.

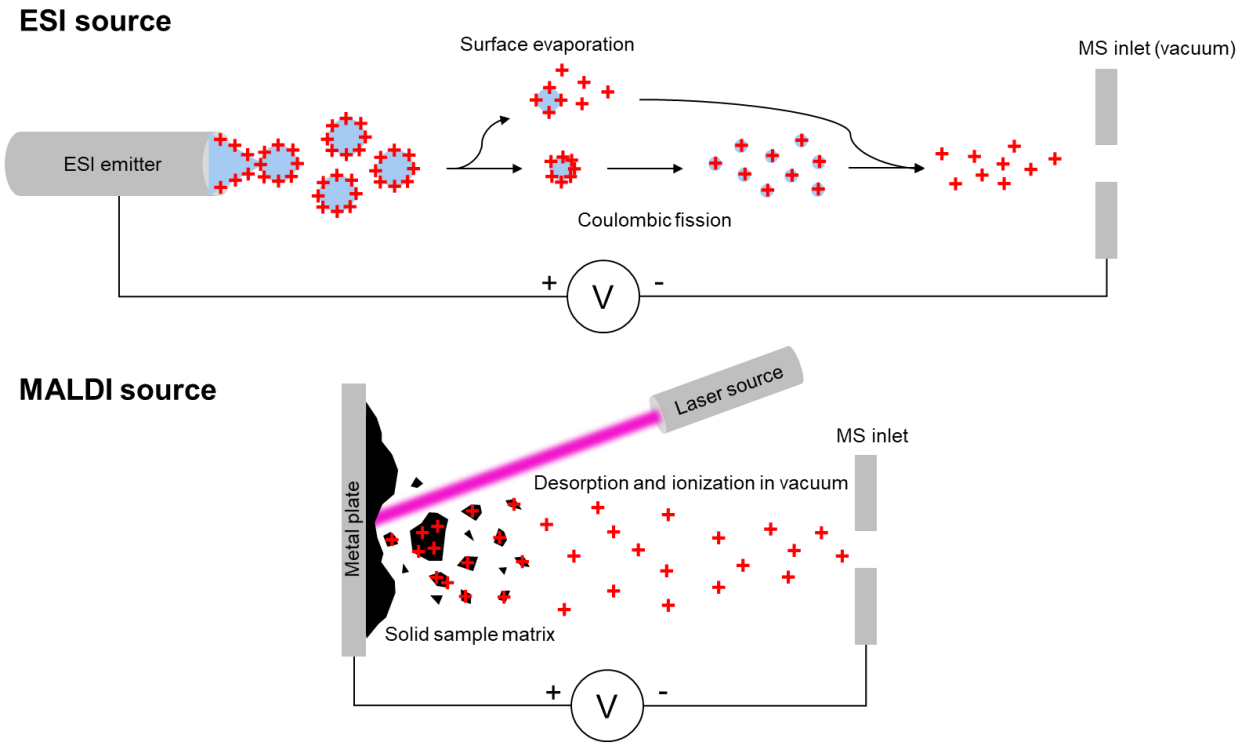


Figure 1.6 LC-MS/MS acquisition modes for untargeted proteomics.

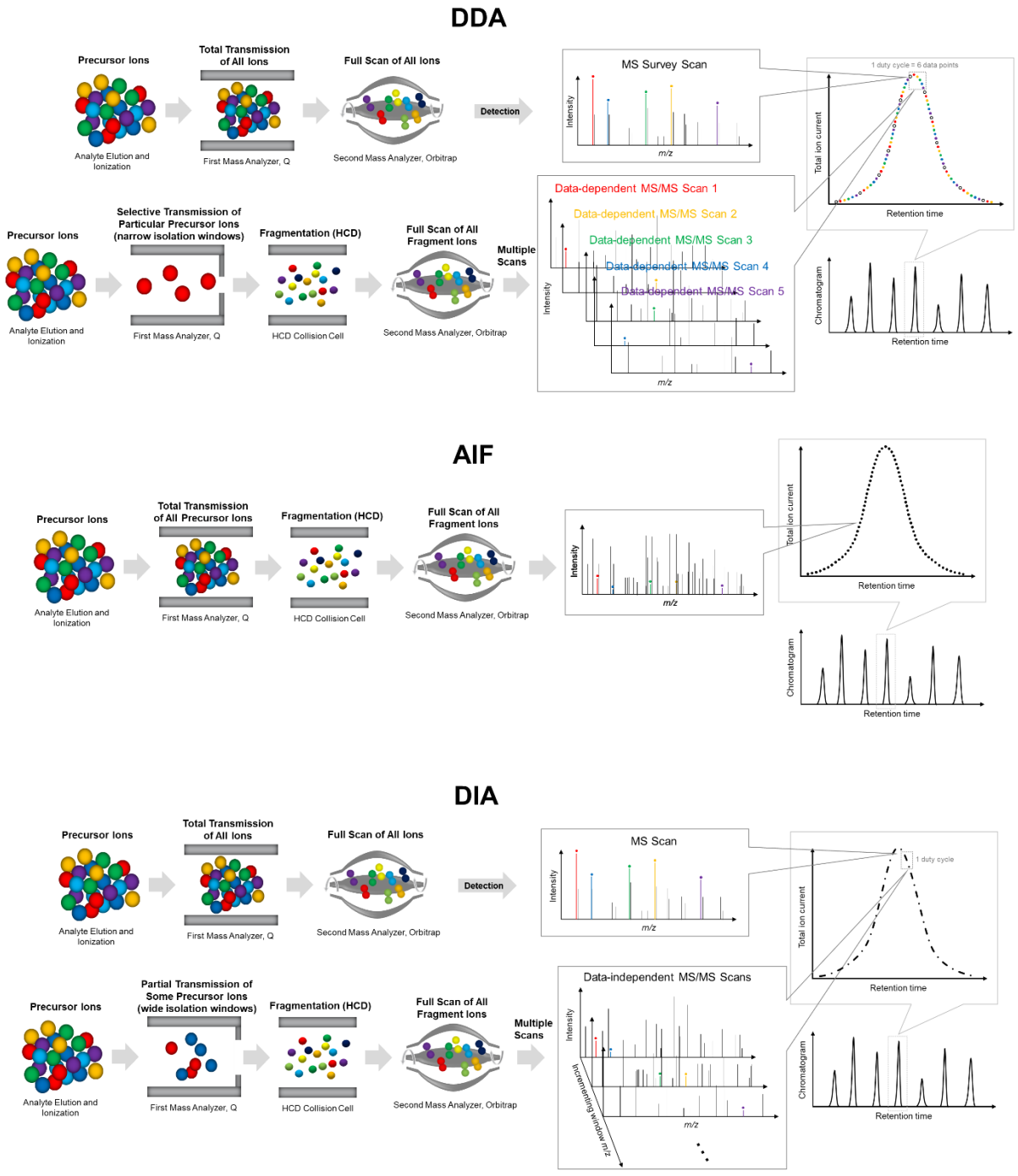


Figure 1.7 LC-MS/MS acquisition modes for targeted proteomics.

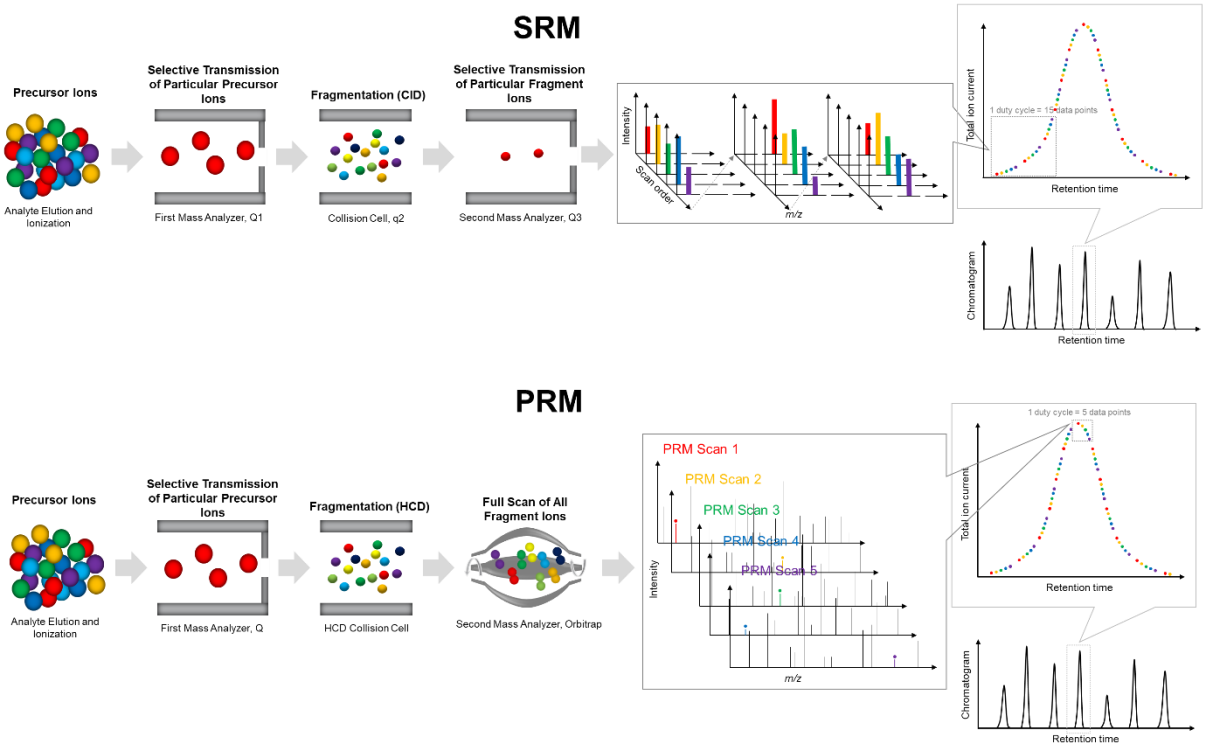


Figure 1.8 Bioinformatics workflow for database search and peptide identification

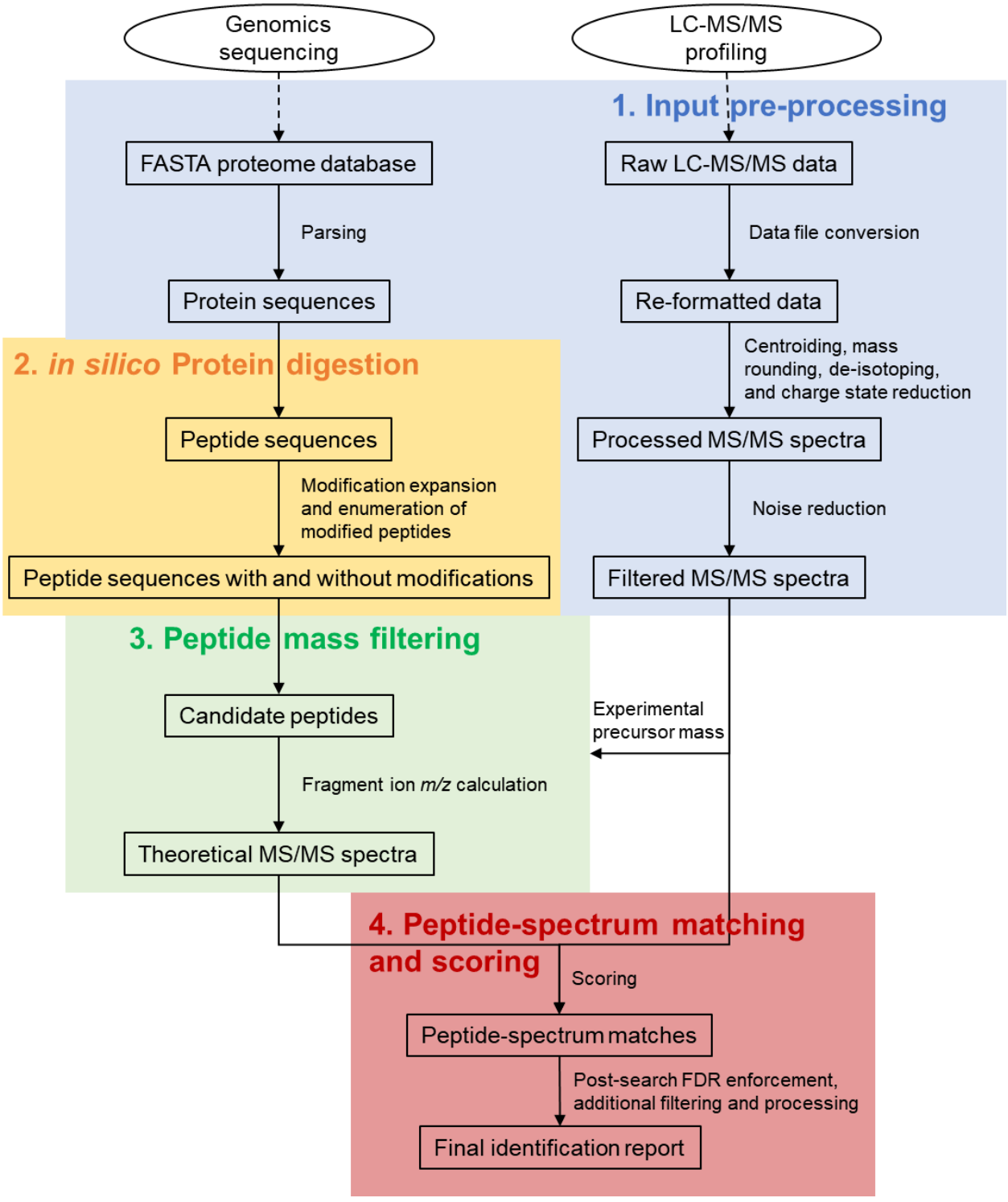


Figure 1.9 Schematics of the compound-centric probe vs. activity-based probe.

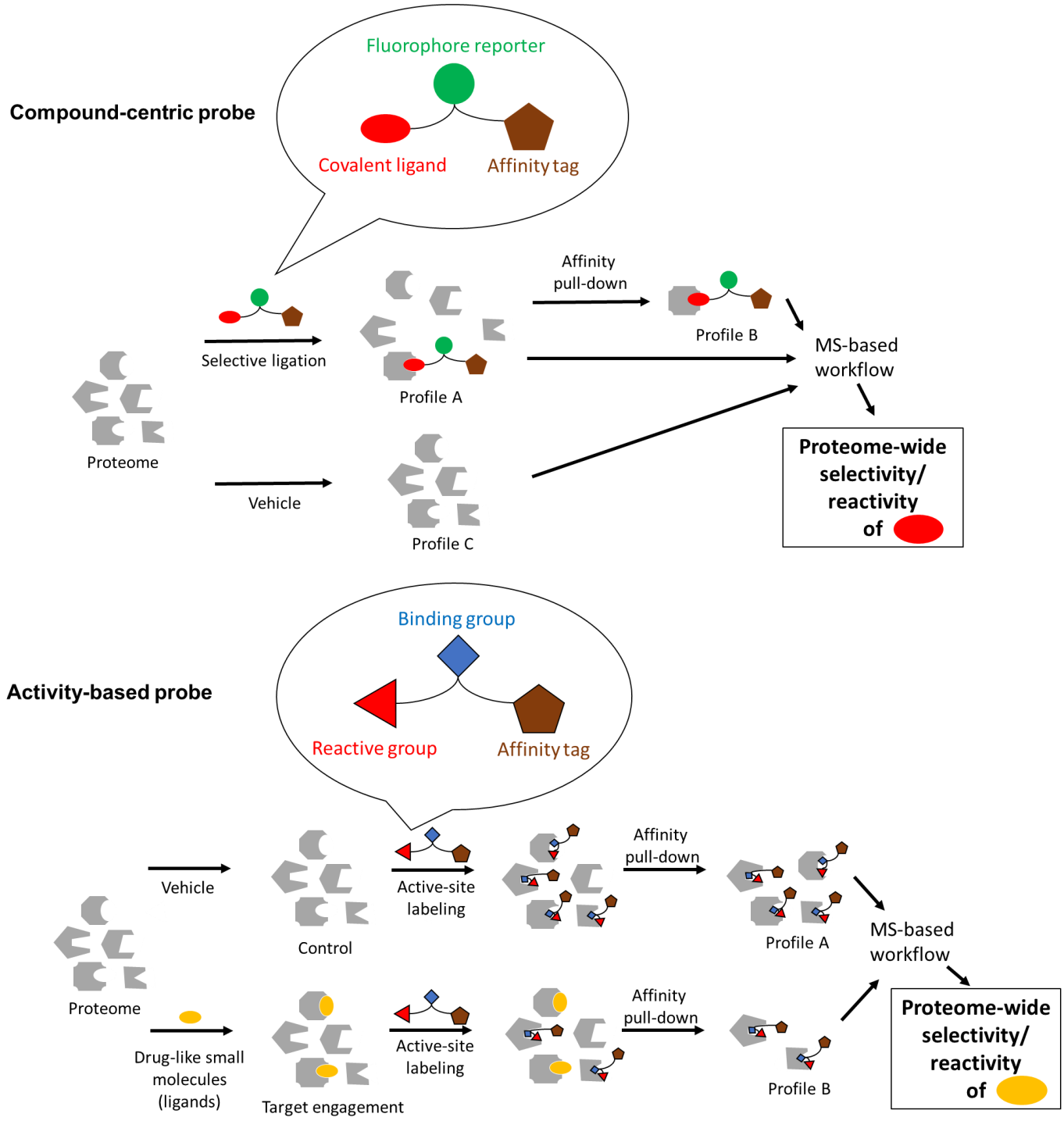
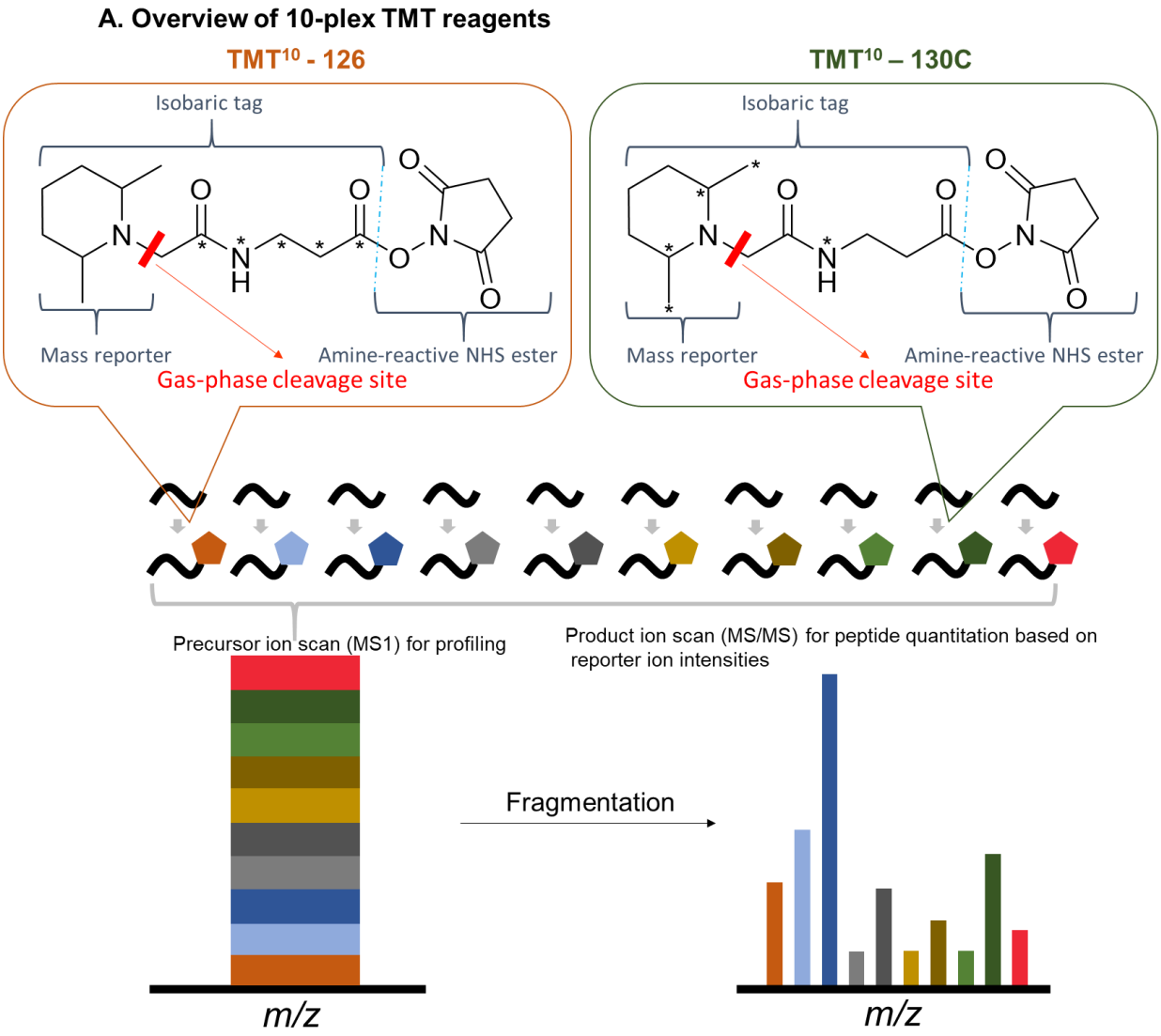


Figure 1.10 TMT enabling accurate and multiplexed target quantitation in chemical proteomics.



B. 10-plex TMT-based chemical proteomics workflow

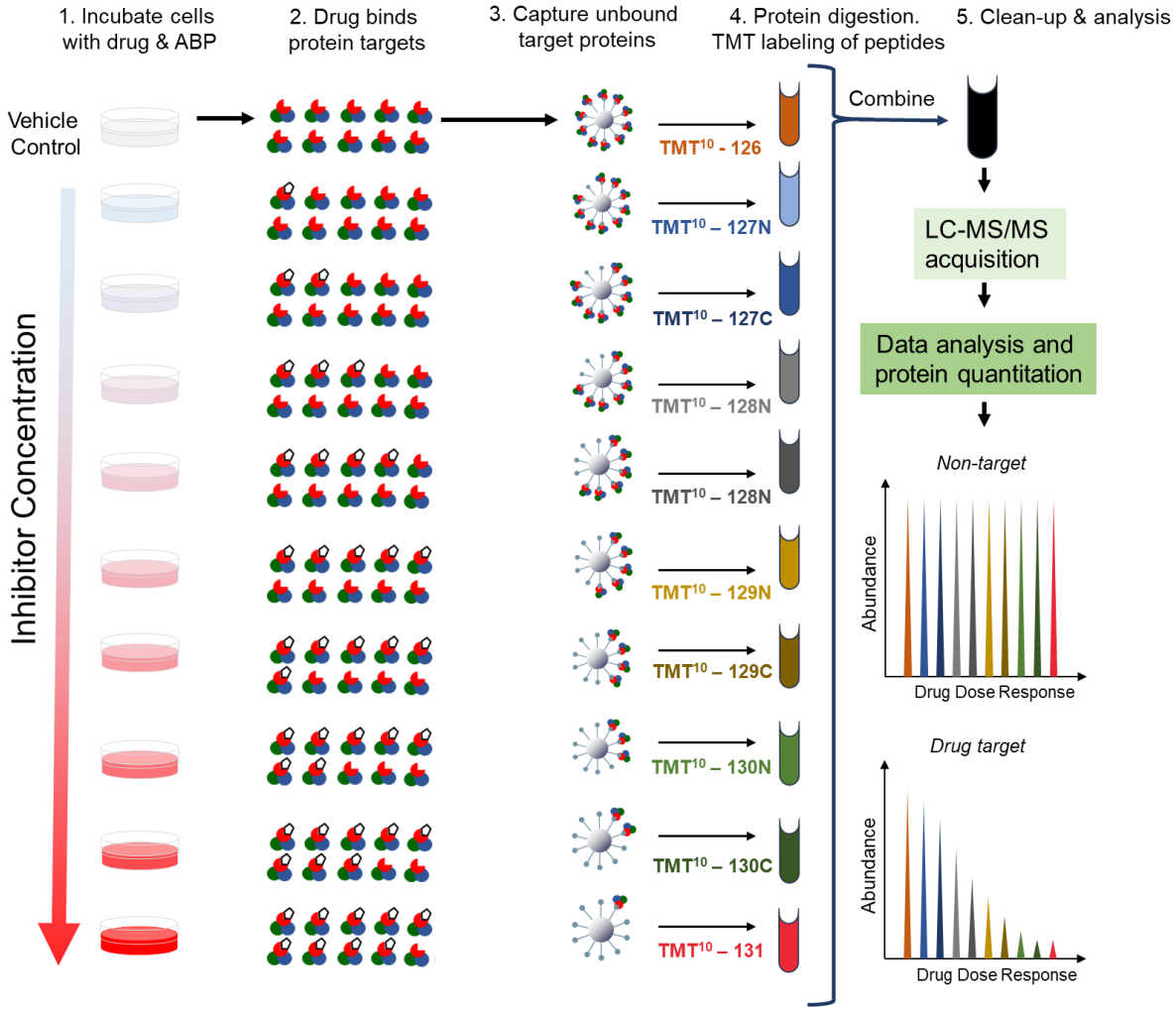
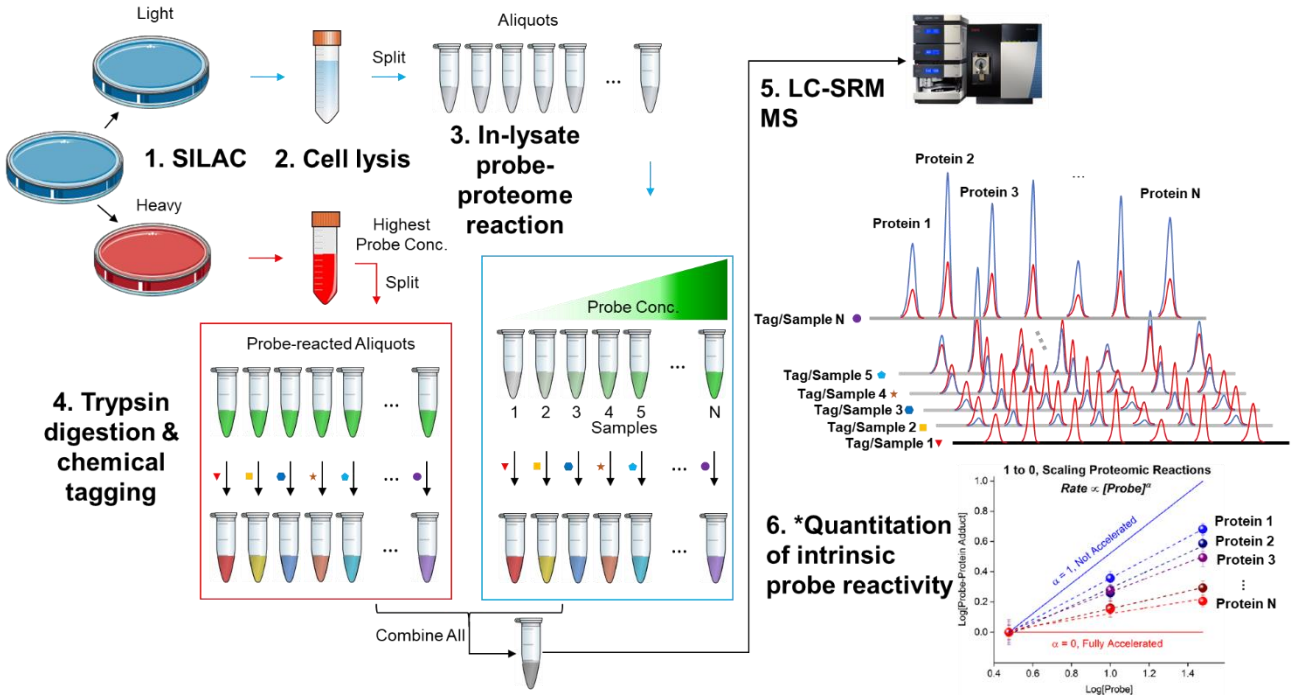
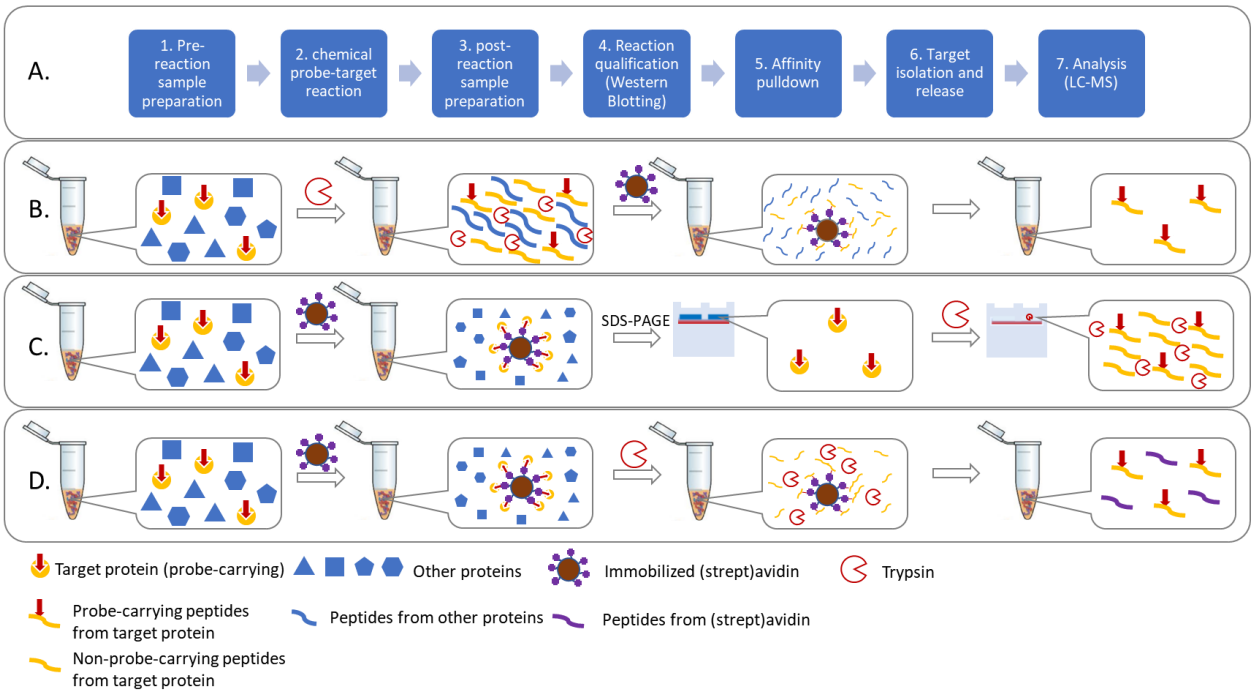


Figure 1.11 Ultra-throughput MRM MS for quantitative scaling of proteome-wide reactivity of activity-based chemical probes.



*Note: plot adapted with permission from (Li, S. et al. Scaling Proteome-Wide Reactions of Activity-Based Probes. Analytical Chemistry 89, 6295-6299). Copyright (2020) American Chemical Society.

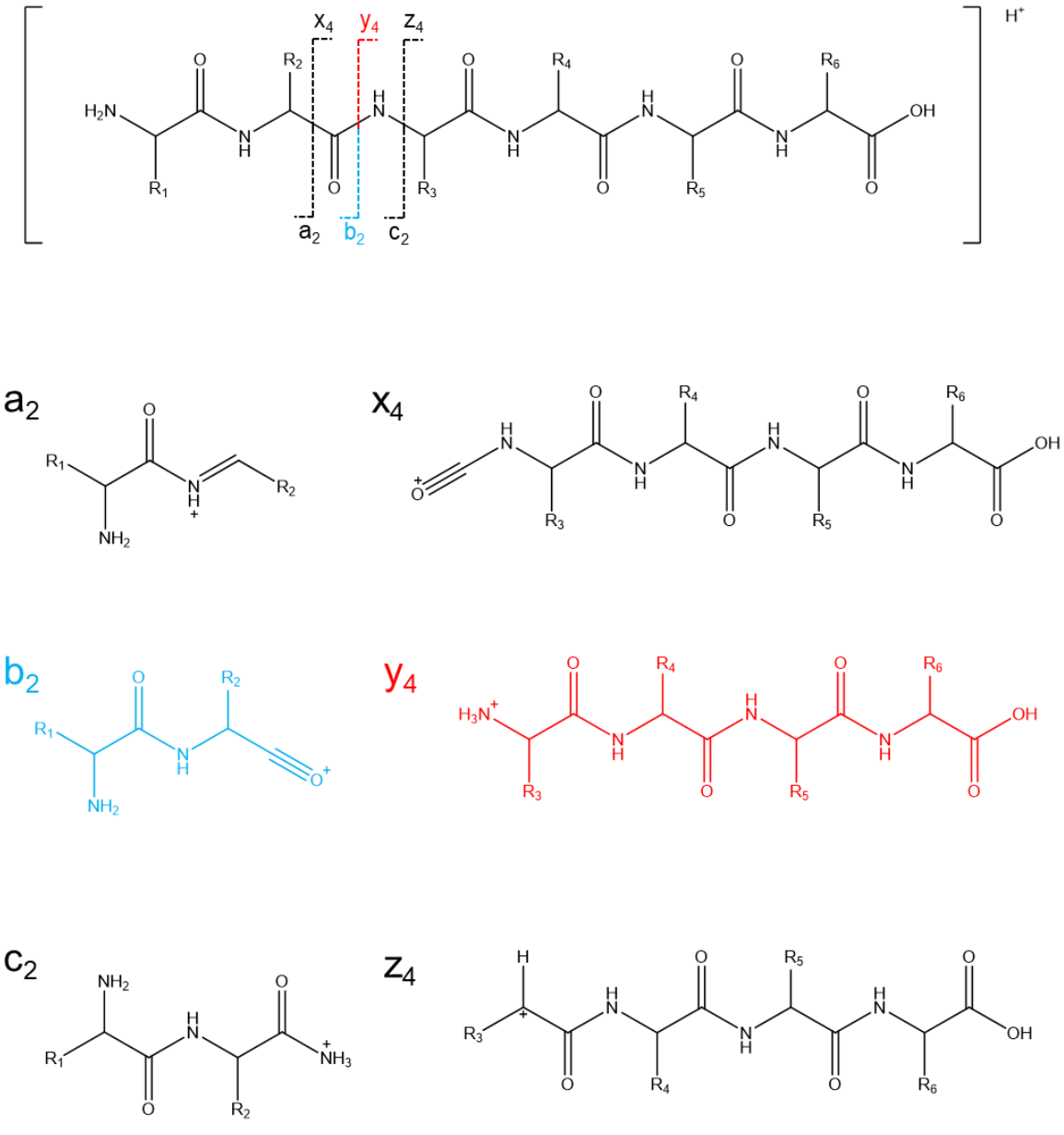
Figure 1.12 Three different affinity capturing approaches used in chemical proteomics.



Note: (A) general workflow chemical proteomics, (B) in-solution trypsin digestion procedure for peptide-level target enrichment, (C) in-gel trypsin digestion procedure for protein-level target enrichment, and (D) on-bead trypsin digestion procedure for protein-level target enrichment

1.6 Chapter 1 Schemes

Scheme 1.1 Six types of sequence ions in gas-phase peptide fragmentation.



1.7 Chapter 1 Tables

Table 1.1 Overview of top 5 most popular database search engines.

Algorithm name	Software package implementation	Scoring system
SEQUEST[127]	Comet[128] (open source), Tide/Crux[129] (open source), Proteome Discoverer (commercial)	• Convert observed and theoretical spectra to frequency domain by fast Fourier transform • Cross-correlation, reporting ratio between zero-offset alignment and nearby alignments
X!Tandem[130]	Parallel Tandem[131] (open source), Trans-Proteomic Pipeline[132] (open source)	• Count matching fragment ions (b and y ions only) on the observed spectrum • Calculate dot product using ion intensities and the number of matching ions • Calculate "hypercore" by multiplying with factorials of the number of assigned b and y ions (hypergeometric distribution) • Build a histogram of scores per spectrum and report its expectation value
MOWSE[133]	Mascot (commercial)	• Calculate the probability (P) if a PSM occurs by chance • Report the score as $-10\log_{10}(P)$
MS-GF+[134]	MS-GF+ (open source), ProteoSAFe (web-based)	• Convert peptide **P** and spectrum **S** into peptide vector **P*** and spectral vector **S*.** • Calculate dot-product Score(**P, S**) = **P*·S***

Andromeda[135]	MaxQuant[136] (freeware)	• Filter top *q* peaks per 100 Da mass interval on the theoretical spectrum • Count matching fragment ions (*k*) on the observed spectrum to all possible theoretical ions (*n*) • Score each mass interval as $-10\log_{10}$ of the probability of matching at least *k* out of the *n* theoretical masses by chance (binomial distribution) • Score the spectrum as the highest mass interval score

Chapter 2 Investigation of Covalent Protein Adducts of 2-Nitroimidazole-ICG as a Hypoxia-targeting Probe in Mouse Tumor

2-Nitroimidazole is a well-known chemical probe targeting hypoxic environments of solid tumors, and its derivatives are widely used as imaging agents to investigate tissue and tumor hypoxia. However, the underlying chemistry for the hypoxia-detection capability of 2-nitroimidazole is still unclear. This chapter reports the deployment of a biotin conjugate of 2-nitroimidazole-indocyanine green (2-nitro-ICG) for the investigation of *in vivo* hypoxia-probing mechanism of 2-nitro-ICG compounds. Mass spectrometry-based proteomics and exhaustive data mining concluded that 2-nitro-ICG and its fragments modified mouse serum albumin as the primary protein target, but at two structurally distinct sites, possibly via two different mechanisms. The identification of probe-modified peptides not only contributes to the understanding of the *in vivo* metabolism of 2-nitroimidazole compounds but also demonstrates a competent analytical workflow that enables the search for peptides with undefined modifications in complex proteome digests.

The *Journal of Mass Spectrometry* published this study under the title "Treasure hunt for peptides with undefined chemical modifications: Proteomics identification of differential albumin adducts of 2‐nitroimidazole‐indocyanine green in hypoxic tumor" as a research article.[137] The reuse of its content in this Chapter 1s permitted by John Wiley and Sons and Copyright Clearance Center under the license number 4816640832152.

2.1 Introduction

2.1.1 Tumor hypoxia and its detection

Hypoxia is a well-renowned vicious low-oxygen condition that that is a salient feature of most tumors. Tumor hypoxia occurs because of a cumulative micro-environmental imbalance between diminished oxygen supply and elevated demand in the region of a fast-growing tumor.[138,139] As a consequence of inefficient local vascular network and hyperactive cellular metabolism, tumor hypoxia is often responsible for severe inhibition of immune cells, strict exclusion of therapeutic agents, and malignant progression of tumors.[140] Thus, detection and measurement of tumor hypoxia can provide invaluable clinical information for initialization and evaluation of anti-cancer therapy.[141-143]

Selectively monitoring an exogenous bio-reducible and traceable marker is one of the mainstream methods for hypoxia assessment, where 2-nitroimidazole derivatives are typically used as the hypoxia tracer for indirect measurements.[144] In contrast to 2-nitroimidazole derivatives as established radiotracers for positron emission tomography (PET) to image tumor hypoxia, indocyanine green (ICG) conjugates of 2-nitroimidazole are a novel class of near-infrared (NIR) fluorescent tumor hypoxia tracers demonstrating superb *in vivo* detectability and biocompatibility.[145-150] Built on a fluorophore analogous to the commercial ICG as the scaffold structure, this class of compounds has evolved for three generations as ICG conjugates differing in linker and imidazole structures, as shown on **Scheme 2.1**.

2.1.2 A brief history of nitroimidazole-ICG probes

On **Scheme 2.1**, all of these nitroimidazole-ICG probes were synthesized at Dr. Michael Smith's laboratory at Department of Chemistry, University of Connecticut, referred to as Generation I[145,146], Generation II[147-149], Generation IIIa[150], Generation IIIb[151], and Generation IIIc (to be used this study) dye. Regarding the lower reduction potential and toxicity of some 4-nitroimidazole derivatives, the initial work focused on the development of 2-nitroimidazole derivatives as chemical probes that target tumor hypoxia, based on the scaffold compound, bis-carboxylic ICG (*compound 1*).[145,148,152-154] In a continuous effort to develop the best probe, the first-generation dye conjugate using an ethanolamine linker (*compound 1-1*)[146], and the second-generation dye conjugate using a piperazine linker (*compound 1-2*)[147-149] were prepared. In both cases, previous *in vivo* studies showed that the dye conjugate was retained in hypoxic tumors, allowing imaging by the NIR fluorescence technique.[145] For the third-generation dye-conjugates, apart from the 2-nitroimidazole-based probes (*compound 2-1 and 1-5-1*), the unsubstituted imidazole derivative (*compound 1-3*) and the 4-nitroimidazole dye-conjugate (*compound 1-4*) were also reported in a recent study.[151]

Despite numerous cases where 2-nitroimidazole derivative-assisted tumor hypoxia detection techniques have been successfully established, a comprehensive understanding of explicit biochemistry for the hypoxia-detection capability of 2-nitroimidazole is still absent. In the early 1980s, Dr. Raleigh and his colleagues demonstrated that misonidazole, a 2-nitroimidazole derivative, could undergo oxidoreductase-catalyzed degradation in a cell-free experimental system.[155] The same research team later illustrated that the 2-nitroimidazole derivative could covalently bind to macromolecules after reductive activation by high-energy photoirradiation.[156] This study also suggested the activated 2-nitroimidazole possessed a higher selectivity for proteins

compared to nucleic acids and the highest reactivity to thiol groups among any other nucleophiles on bovine serum albumin in solution. Over the years, a predominant number of *in vitro* and *in vivo* studies on mammalian cells have been dedicated to analyzing 2-nitroimidazole's metabolites.[157-160] According to some recent studies, the hypoxia-detection capability of 2-nitroimidazole-mediated imaging methods mostly arose from the regional accumulation of specific 2-nitroimidazole metabolites, such as the glutathione conjugate of 2-aminoimidazole.[158,159] However, macromolecular adducts of 2-nitroimidazole have not been thoroughly examined since the initial study in the 1990s.[156,161]

2.1.3 MS-based proteomics to identify 2-nitroimidazole targets

To unleash the full potential of 2-nitroimidazole derivatives and similar compounds for targeting tumor hypoxia at a higher efficiency, accuracy, and biocompatibility, it has become imperative to re-examine the reactivity of 2-nitroimidazoles from the proteomics perspective. Related questions (**Scheme 2.2a**) are: (1) which proteins 2-nitroimidazole-indocyanine green (2-nitro-ICG) probes preferentially bind to; (2) how these hypoxia-driven probe-protein reactions initialize, progress, and terminate; and (3) to what extent these reactions can perturbate the proteome of tumor cells *in vivo*. Therefore, a biotinylated version of the 2-nitro-ICG probe (biotin dye, **Scheme 2.2b**) was developed and applied to facilitate a modern mass spectrometry-based proteomics investigation (**Scheme 2.2c**) of the chemistry for *in vivo* hypoxia-probing mechanism.

Nonetheless, given the large molecular weight of 2-nitro-ICG probes, target analytes (proteins modified by 2-nitro-ICG *in vivo* and digested by trypsin) are potentially labile while their affinity enrichment efficiency may be low. Substantial degradation of probe-protein adducts may occur during sample preparation. It is usually prudent to project a data utility loss of 80% to ensure sufficient analysis of the resulting highly convoluted proteomics profiling data. In an attempt to

accomplish an ambitious objective aiming at the identification of modification sites, the investigator is likely to experience extraordinary challenges when both the target proteins and their measurable modification (regarding the increase in mass of the peptide after its modification by the chemical probe) are undefined. Consequently, the data often remains uninterpretable for identifying probe-modified peptide via the typical data processing workflow used in a shotgun proteomics experiment.[162,163] While the direct identification of covalent adducts appears unachievable, the "gray area" of the profiling data, where a large volume of valuable information on modified peptides deposits as unassigned spectra, can offer a second chance.[164] As the task of "rescuing" spectra featuring probe-modified peptides is extraordinarily labor-intensive and frustrating, methods to alleviate such arduousness for the identification of undefined covalent adducts are in immediate demand.

Herein, this study introduces a front-to-end analytical workflow that dictates the successful discovery of a primary protein target and its modification sites for a 2-nitroimidazole-based *in vivo* chemical probe. In this workflow, probe-reacted proteins are detected as target analytes using fluorescence imaging and isolated using avidin affinity pull-down in conjunction with multiple techniques of gel electrophoresis. These electrophoresis techniques include the sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), isoelectric focusing-polyacrylamide gel electrophoresis (IEF-PAGE), and a combination of the two as the two-dimensional gel electrophoresis (2-DE). The isolated target analytes are measured as tryptic digests using LC-MS/MS. The profiling data is processed and interpreted using an approach of product-ion-oriented data mining for the identification of the target protein and its modification sites. In general, this approach bypasses the search restriction on peptidyl precursor ions at MS level, weights the detection of surrogate peptidyl product ions and probe fragment ions at MS/MS level, bridges the

gap between MS and MS/MS data with chromatographic information, reconciles MS data for evidence supporting the assignment of peptide identities, thus distinguishes otherwise unassignable spectra featuring probe-modified peptides. This work broadens the perspective on the hypoxia-probing mechanism of 2-nitroimidazole compounds. More significantly, it conveys a practical troubleshooting guideline for similar chemical proteomics studies routinely conducted during the research and development of new active pharmaceutical ingredients, where the measurable chemical modifications on target and off-target proteins are usually unknown or mutable.[165]

2.2 Experimental

2.2.1 Materials

Solvents, reagents, stock buffer solutions, and equipment were purchased from Bio-Rad Laboratories Inc (Los Angeles, CA), Fisher Scientific Co LLC (Hanover Park, IL), Life Technologies (Grand Island, NY), Sigma-Aldrich Inc (St. Louis, MO), and Thermo Fisher Scientific LLC (Asheville, NC). Details of materials, mouse tumor model preparation, and fluorescence imaging are included in the supplemental information.

2.2.2 Tumor lysis and protein extraction

The excised solid mouse tumors had approximate dimensions of 1 cm x 1.5 cm x 0.2 cm. For each tumor sample, the solid excision was homogenized in a 7-mL Dounce tissue homogenizer (Bellco Glass, Vineland, NJ) with 1 mL of a working lysis buffer (WLB) on ice until visually homogeneous. A 5-mL sterile disposable syringe with a 20-gauge needle (BD, Franklin Lakes, NJ) was used to assist the disruption of cellular membranes. The WLB contained 1% of Halt™ EDTA-free 100X protease inhibitor cocktail in Pierce™ IP Lysis Buffer. The homogeneous suspension was split and transferred to 1.5-mL microcentrifuge tubes. The collection tube of the homogenizer was washed with 200 μL of WLB, which was combined with the suspension. For maximal extraction of the cellular protein content, the suspension underwent two freeze-thaw cycles (frozen at -80 °C and thawed on ice), three rounds of centrifugation and resuspension. In each round, the homogeneous tumor suspension was centrifuged for 1 h at 16100 rcf, 4 °C. In total, an additional 1 mL of WLB was used to resuspend the pellets. All the supernatants were collected and combined as a whole crude lysate for each tumor sample. Both the crude lysate and pellets

were saved for further analyses. Next, the crude lysate was purified with Zeba™ Spin Desalting Columns to remove its small-molecule content by following the manufacturer's instruction. For each sample, approximately 2 mL of purified lysate was collected as a protein extract. Each extract was quantified via the standard BCA assay protocol and aliquoted into new 1.5-mL microcentrifuge tubes, which were stored at -80 °C before further analyses. The final protein extracts had a concentration of approximately 7 mg/mL.

2.2.3 Affinity enrichment of covalent adduct

For each avidin affinity pull-down experiment, a 100-µL aliquot of protein extract was diluted for a protein concentration of 2 mg/mL with freshly-prepared WLB and mixed with an equal volume solution of 8 M urea in WLB, resulting in a pre-binding solution (PS) containing 1 mg/ml proteins and 4 M urea. Afterward, the PS was mixed with either Pierce™ NeutrAvidin™ agarose resins or Pierce™ high-capacity streptavidin agarose resins (Thermo Fisher Scientific) according to the proper the protein amount vs. binding capacity ratios. After overnight incubation at 4 °C with constant agitation on the HulaMixer, appropriate washing steps were performed based on the manufacturer's instruction manual. Briefly, the adduct-bound resins were washed with three different washing buffers, which were prepared in situ. Washing buffer 1 (WB1) was prepared as 4 M urea in Pierce™ IP lysis buffer. Washing buffer 2 (WB2) was prepared as 5% isopropanol in PBS (pH 7.4). Washing buffer 3 (WB3) was prepared as a solution of 20 mM of Tris-HCl, 50 mM of NaCl, whose pH was adjusted to 8.0. The adduct-bound resins were washed three times of WB1 and WB2, twice of WB3, with either 20X volume of the settled resins for the high-capacity streptavidin-agarose or 10X volume of the settled resins for the NeutrAvidin-agarose. After resuspension of resins in the washing buffers via brief vortexing, centrifugation was performed for 5 min at 2000 rcf, 4 °C, to pellet the resins. Supernatants were collected as washes and stored at -

20 °C for checkpoint analyses. For releasing resin-bound adducts, an equal resin-volume of a freshly prepared working elution buffer (WEB) was added to each 1.5-mL microcentrifuge tube that contained washed adduct-bound resins. The WEB was prepared by mixing 2-mercaptoethanol, Laemmli sample loading buffer, and a stock solution of biotin in WB3. The resulting resin suspension included 2.5% of 2-mercaptoethanol, 1X Laemmli sample loading buffer, and 4 mM of biotin, considering the volume of resins. The microcentrifuge tube was then heated in a hot water bath at 95 °C for 15 min and cooled down to room temperature, protected from the light. The heated resin suspension (HRS) was either used immediately or temporarily stored at 4 °C for either analytical or preparative gel electrophoresis experiments.

2.2.4 Gel electrophoresis and fluorescence detection

For the analytical SDS-PAGE, a small portion of each sample (5 to 20 µL of either protein extract, washes from affinity pull-down experiments, the supernatant of HRS, or whole HRS) was loaded onto a Mini-PROTEAN® TGX™ precast gel (1.0-mm, 10-well, Bio-Rad). For the preparative SDS-PAGE, approximately 90 µL of whole HRS was loaded into each well of an in-house prepared gel (1.5-mm, 5-well). The gel consisted of a top stacking layer and a bottom resolving layer with a volume ratio of 2:8. The stacking layer was cast with 4% acrylamide/bis-acrylamide in a buffer of 0.5 M Tris-HCl and 0.4% SDS, whose pH was adjusted to 6.8. Whereas, the resolving layer was cast with 12% acrylamide/bis-acrylamide in a buffer of 0.5 M of Tris-HCl and 0.4% SDS, whose pH was adjusted to 8.8. Together with the samples, 5 µL of either SeeBlue™ Plus2 pre-stained protein standard (Life Technologies) or Precision Plus Protein™ dual color standards (Bio-Rad) was loaded to the gel as molecular weight reference protein standards. The subsequent electrophoresis was performed either at constant 120 V for 90 min or at a two-step

setting of 120 V for 40 min and 200 V for 20 min until complete migration of the fronting band to the bottom of the gel.

For the implementation of preparative 2-DE, we adopted a high-compatibility but low-cost protocol.[166] This protocol seamlessly merged two distinct gel electrophoresis techniques: vertical isoelectric focusing-polyacrylamide gel electrophoresis (vIEF-PAGE) and SDS-PAGE, with minimal instrumentation requirements. As the first part of this 2-DE procedure, vIEF-PAGE was performed according to its manufacturer's instruction. Briefly, 20 µL of protein extract was mixed with 20 µL of provided IEF sample buffer at room temperature. The resulting 40 µL sample solution was loaded to a provided precast IEF gel in a cassette assembly accommodated in the electrophoresis cell. The three-step electrophoresis was operated at 50 V for 60 min, 200 V for 60 min, and 500 V for 30 min. Afterward, the IEF gel was incubated in 20% TCA for visualization of white bands via protein precipitation and sliced into strips along the edge of each well, parallel to the well-dividers, with a piece of thin glass on a clean glass surface. The gel strips were dehydrated in acetonitrile until turning opaquely white and rehydrated in a solution of 2.5% 2-mercaptoethanol in the stacking gel buffer. As the second part of this 2-DE procedure, the prepared strip was loaded onto an in-house-cast gel (1.5-mm, single-well, 10% polyacrylamide). The following SDS-PAGE was performed in the same manner as previously described.

After electrophoresis, the protein bands were fixed with a solution of 10% acetic acid and 50% methanol in water. The gels were imaged with an LI-COR Odyssey imaging system (LI-COR, Inc. Lincoln, NE) with setting as a dual channel (700 nm and 800 nm; the 700-nm channel was monitored for visualizing the reference protein ladder), 0.75 mm focus offset, 7.0 intensity, and highest quality. After successful detection of the fluorescent protein band, the gels were stained with QC Colloidal Coomassie Stain (Bio-Rad), placed in a clear resealable plastic pouch,

and imaged with a document scanner. Finally, the target protein band or spot was excised and diced into 1 mm x 1 mm cubes with a sterile scalpel on a clean glass surface. Gel cubes were transferred to clean 1.5-mL microcentrifuge tubes either for immediate preparation for in-gel digestion or temporary storage at 4 °C.

2.2.5 In-gel trypsin digestion

The excised gel bands or spots were washed with acetonitrile for destaining, treated with DTE for reduction, IAA for alkylation, and incubated overnight with trypsin for digestion. Refer to Text S6 for details. Afterward, peptides were extracted with an appropriate amount of acetonitrile from the post-digestion gel dices, concentrated in SpeedVac, and acidified with formic acid (to $pH \leq 3$) for desalting with the solid-phase extraction (SPE) technique. For desalting peptide samples, Pierce™ C18 StageTips (Thermo Fisher Scientific) were used. The SPE procedure was performed according to the manufacturer's instructions. The desalted tryptic digests were concentrated in the SpeedVac, lyophilized, and reconstituted to 10 µL.

2.2.6 LC-MS/MS analysis

The in-gel tryptic digests were analyzed on a Q Exactive HF mass spectrometer equipped with the nanospray ionization (NSI) source coupled with an UltiMate 3000 UHPLC system (Thermo Fisher Scientific).

For the LC part, a nanoEase *M/Z* Peptide BEH C18 column (25 cm length, 75 µm diameter, 1.7 µm particle size, and 130Å pore size) was used for peptide separation. The autosampler temperature was 4.0 °C. Column oven temperature was 50.0 °C. Sample injection volume was 1.0 µL. Mobile phase flow rate was set as 300 nL/min. Solvent A was 0.1 % formic acid in water, and solvent B was 0.1 % formic acid in acetonitrile. The complete 90-min LC method consisted of a

10-min sample-loading period, a 50-min mobile phase linear gradient period, and a 30-min post-gradient column flushing and equilibration period. Specifically, the method profile (% for Solvent B at runtime) was 4% at 0 to 10 min, 30% at 50 min, 90% at 60 to 70 min, and 4% at 72 to 90 min.

For the mass spectrometer part, Xcalibur v2.8 software controlled the instrument. The mass spectrometer operated in DDA mode for monitoring positive ions at a spray voltage of 1500 V. For MS1, the mass range was from 300 to 1800 *m/z*. The Orbitrap mass analyzer was set with a resolution of 60,000, an AGC target of 1e6, and a maximum ion time of 60 ms. For data dependent MS2, the quadrupole was set with an isolation window of 2.0 *m/z*. The Orbitrap was set with a resolution of 15,000, an AGC target of 1e5, and a maximum ion time of 40 ms. This DDA method allowed up to 20 MS/MS scans per duty cycle, and a stepped normalized collision energy (NCE) of 27. Precursors that triggered MS/MS scans were dynamically excluded from repetitive MS/MS scans for 40 s. Charge state exclusion was enabled to reject precursor ions with charge states beyond the range of +2 to +8. Peptide match option was set at preferred. MS/MS spectra were collected as the centroid data type.

2.2.7 LC-MS/MS data processing

For peptide and protein identification, the mass spectral data were searched against a mouse reference proteome database (Swiss-Prot, Mus musculus, UP000000589, last modified on October 22, 2018) containing 16,997 mouse proteins. Database searches were performed with both MaxQuant (version 1.6.1.0) and MODa (version 1.60).

For MaxQuant-based searches, essential parameters were set as follows: 1% peptide-level false discovery rate (FDR), 1% protein-level FDR, 1% modification site FDR, a minimum peptide length of 5, a minimum score of 0 for unmodified peptides, a minimum score of 10 for modified peptides, a minimum unique peptide number of 0, a minimum razor peptide number of 1, an

MS/MS mass error tolerance of 20 ppm, a peptide length range of 8 to 25 for unspecific search, a maximum missed peptide cleavage of 3, a maximum peptide mass of 8000 Da, and a revert decoy mode. For the setting of modification inclusion list, oxidation on methionine, acetylation on protein N-terminus were set as variable modification. Besides, carbamidomethylation, biotinyl piperazine-2-nitroimidazole-ICG (original form, +1284.5250 Da), biotinyl piperazine-2-aminoimidazole-ICG (reduced form 1, +1254.5519 Da), or biotinyl piperazine-2-aminoimidazole-ICG (reduced form 2, + 1239.5410 Da) on cysteine was set as a variable modification. Up to 5 variable modifications per peptide were allowed.

MODa-based searches were performed in single-blind mode (maximum one modification per peptide). Key parameters are two enzymatic termini, no missed cleavage, a modification mass range of 150 to 1500, no fixed modification, a fragment ion tolerance of 0.01 Da, no precursor ion auto-correction, and a precursor mass tolerance of 2 Da. FDR of 1% was enforced for separate searches against the database appended with reverted decoy protein sequences. Additional data processing was performed with the assistance of ProteoWizard Toolkit[167] for conversion of data format, visualization, and extraction of ion chromatograms and spectral binary datasets from the raw data files.

2.2.8 Mouse serum albumin modeling and molecular docking

The sequence of mouse serum albumin (AlbM, P07724) was obtained from UniProt protein data repository in FASTA format.[168] The AlbM structural model was built using the SWISS-MODEL (https://swissmodel.expasy.org), based on target-template sequence alignment and homology.[169] After the template database query with the sequence of AlbM as the input, all 42 available templates were sorted by their global model quality estimation (GMQE) scores. With a GMQE score of 0.86 being the highest among 42 available serum albumin templates, the X-ray

crystal structure of rabbit serum albumin (PDB accession: 3V09)[170] was selected as the template to generate the three-dimensional (3D) structure of AlbM in PDB format.

The virtual probe-protein binding (docking) study was performed using AutoDock Vina[171] and AutoDock Tools version 1.5.6[172]. For docking model preparation, the 3D structure of AlbM, water molecules were removed, and polar hydrogen atoms were added. The 3D structure of biotin dye was constructed in two-dimensional MOL format using ChemDraw Prime 16.0.1.4 and converted to three-dimensional MOL2 format using Avogadro[173]. The following 3D structure of biotin dye was prepared for docking by assigning torsional bonds. For peptide LPCVEDYLSAILNR, the grid box with a size of 60 × 30 × 36 Å was allocated at the center of the binding cavity with x, y, and z coordinates of 30, 15, and 12. For peptide DTCFSTEGPNLVTR, the grid box with a size of 30 × 30 × 30 Å was centered on the cysteine residue with x, y, and z coordinates of 32, 60, and 20. Other parameters were set as default values. The docked models with the highest affinity (-8.5 kcal/mol for LPCVEDYLSAILNR and -6.1 kcal/mol for DTCFSTEGPNLVTR) were selected result to demonstrate the probe-protein interaction before the formation of covalent adducts.

2.3 Results and Discussion

2.3.1 The biotinyl 2-nitroimidazole-ICG

By conjugating a biotin moiety to the previously reported 2-nitro-ICG chemical probe,[148,150] we designed biotinyl 2-nitroimidazole-ICG (referred to as the biotin dye, **Scheme 2.2b**) as a "dual-functional" chemical probe. In comparison, most chemical probes used in proteomics contain one of two groups for detection: a fluorescence reporter group or biotin "affinity handle.[174] Ideally, the "dual-functional" design of the biotin dye would offer multiple analytical advantages including flexible and confident detection, leveraged sensitivity, and surface chemistry-compatibility.

With its exceptional affinity to avidin ($K_d \sim 10^{-15}$ M), the biotin moiety is routinely used as a propagable "affinity handle" on the chemical probe for flexible detection and enrichment of probe-reacted proteins. The downstream sample preparation is supported by a broad range of commercially available products for avidin-based immunoprecipitation and detection.[175] By design, biotin dye allows solid-phase affinity pull-down of probe-reacted proteins from complex tumor tissues. Besides, the biotin moiety on the adducts can be detected by Western Blot analysis (**Figure 2.1**), offering a complementary detection approach to the fluorescence based on the ICG fluorophore of the probe-modified proteins.

The optical absorption and emission peaks, as well as the extinction coefficient of the biotin dye, were measured (**Table 2.1**) by the Zhu group at Department of Biomedical Engineering, University of Connecticut. The absorption and fluorescence emission peaks were similar to those of previously reported ICG probes, while its extinction coefficient was about half of that of the

"fully-loaded" 2-nitro-ICG probe and two-thirds of that of the "half-loaded" 2-nitro-ICG probe, in agreement with the reduced conjugation system in the biotin dye.

2.3.2 Selective protein modification by biotinyl 2-nitroimidazole-ICG in hypoxic mouse tumor

In vivo fluorescence intensity for tumor peaked between minute 5 to 15 post-injection and declined rapidly afterward (**Figure 2.2a**). The average tumor fluorescence intensity in a range of 1 to 60 minutes for the biotin dye was the lowest compared with other ICG probes (**Figure 2.3**). This observation was likely due to the optical property of the probe and its diffusion physics in solution and intercellular translocation within the tumor tissue.[176] After 60 minutes, the fluorescence signal intensity from the biotin dye converged towards that of the "half-loaded" 2-nitro-ICG probe, yet remained weaker than that of the "fully-loaded" 2-nitro-ICG probe (**Figure 2.3** and **Figure 2.4**), which might be explained by the halved stoichiometry of the hypoxia targeting 2-nitroimidazole functional group.

The biotin dye remained highly fluorescent in either the free molecule or adduct forms at all three analytical levels (**Figure 2.2**). At the *in vivo* level, fluorescence images of the probe-injected mouse indicated that the dye molecules circulated throughout the cardiovascular system of the mouse after a short period upon the injection. At the *ex vivo* level, only the tumor, kidneys, and liver exhibited fluorescence in the specified optical condition 48 hours post-injection. As expected, the majority of dye molecules underwent fast excretion via the renal route while a smaller population of them sustained delayed elimination and probable enzymatic degradation through the biliary and hepatic routes.[154,177] It is known that 2-nitroimidazole derivatives are reactive and predominantly retainable under hypoxic conditions in cells,[144] which is consistent with the tumor fluorescence images. Accordingly, the biotin dye successfully reproduced an

imaging profile of *in vivo* and *ex vivo* fluorescence akin to those previously reported for original 2-nitro-ICG probes.[148,150]

At the tumor lysate level, proteins separated by SDS-PAGE were analyzed for fluorescence emission at 800 nm. Notably, there was only a single protein band with a molecular weight of about 70 kDa (left, **Figure 2.2b**) showing intense fluorescence signal, despite the high complexity of its sample matrix as a whole lysate of the tumor (right, **Figure 2.2b**). Furthermore, this band remained intense upon Coomassie Blue staining, compared to intensities of other bands in the entire lane (right, **Figure 2.2b**). These observations suggested that (1) the biotin dye was attached to one protein as the primary target, or less likely a few co-migrating proteins, (2) the probe-protein adducts were covalent because the denaturing condition of SDS-PAGE did not dissociate the adducts, and (3) the protein target would probably be of high abundance.

2.3.3 Identification of protein targets for modification by biotinyl 2-nitroimidazole-ICG

To identify the probe-modified protein, we performed proteomics profiling for two preparations of tumor lysates upon treatment with the biotin dye. One sample (Sample 1) was from the lysate of fluorescent tumor tissue. The lysate was enriched using immobilized avidin and then resolved by SDS-PAGE. The other sample (Sample 2) was from 2-DE-resolved proteins without the affinity enrichment. LC-MS/MS profiling data of peptides from in-gel digestion of both samples were searched against the mouse proteome database (Swiss-Prot, Mus musculus, UP000000589, last modified on October 22, 2018), using Andromeda of MaxQuant.[163] The top hit for both samples was AlbM with 47 peptides identified for the Sample 1 and 55 for Sample 2, among which 34 peptides were shared for both samples (**Table 2.2**). Co-migrating proteins were also identified. In total, MaxQuant search identified 376 peptides corresponding to 75 proteins for

Sample 1 and 377 peptides corresponding to 155 proteins for Sample 2. Among these identified proteins, only 29 proteins (**Table 2.3**) were shared by both samples. No oxidoreductase was on the identified list of proteins. Oxidoreductases play an indispensable role as the bio-reductive activator in the metabolic pathway of 2-nitroimidazole, a prerequisite for the detection of hypoxia.[155,178] This observation was not surprising. The sample preparation workflow of this study targeted proteins with detectable fluorescence, and the number of fluorescence adducts of reduced forms of the probe and oxidoreductases could be too low to call for proteomics profiling.

The inclusion of the theoretical masses of both biotin dye and its anticipated reduced form (biotinyl 2-aminoimidazole-ICG)[156] as variable modifications on cysteine for the database search did not identify peptides with either modification (**Scheme 2.3**). Presumably, the reducing environment of hypoxic tumor favors the reduction of 2-nitroimidazole to 2-aminoimidazole that has increased electrophilicity for forming covalent adducts with nucleophiles on the protein.[156] Since Andromeda of MaxQuant is a commonly used restrictive search engine for proteomics profiling,[163] MaxQuant requires an input of specified searching constraints for expected modifications to be either fixed or variable.

2.3.4 A general data analysis workflow for identifying peptides with defined and mutable modifications

To identify probe-carrying peptides, we performed an alternative database search using MODa[179] (**Scheme 2.3**). MODa is an unrestrictive or "blind" search algorithm that is based on the alignment of peptide sequence tags. It was developed for discovering unknown protein post-translational modifications (PTMs) without any prerequisite knowledge on modification targets or modification-induced peptide mass shifts.[179] Briefly, MODa first performs *in silico* digestion, which converts protein sequences from a proteome database into tryptic peptide sequences, and

calculates theoretical *m/z* values for b and y ions of each peptide to construct two theoretical MS/MS spectra, respectively. In contrast, restricted database search engines, like the Andromeda of MaxQuant, enumerate all possible fragment ions and combine their *m/z* values altogether to generate complex theoretical MS/MS spectra for in-depth statistical analyses and comparisons to experimental MS/MS spectra.[163] Second, MODa creates short sequence tags of 3 to 4 amino acid residues in length, using consecutive theoretical b or y ions, for all the *in silico* peptides. Third, MODa compares the resulting short tags with ion patterns in observed MS/MS spectra. Based on the Needleman–Wunsch algorithm,[180] MODa scores peptide-spectrum matches (PSMs) according to numbers of sequence-tag matches, mismatches, and gaps, and assigns up to five candidate peptides with highest scores to each experimental MS/MS spectrum. The precursor mass value of the spectrum is then compared with calculated masses for the candidate peptides to examine for any possible mass shift that is potentially attributed to a peptide modification, followed by further verification based on sequence-tag gaps in the previous analysis. When there is no mass shift, the peptide is assigned as a native one. Finally, to improve the specificity of the peptide identification, a pre-set FDR threshold is ready to be implemented if decoy protein sequences are appended to the input proteome database. For the FDR-based filtering, MODa examines the statistical distributions of sequence tag-matching scores assigned to co-identified decoy peptides and actual peptides, respectively, adjusts the score cut-off for the peptide candidate output according to their score distributions, and eliminates a large number of low-quality peptide candidates. In contrast to the original candidate pool, the FDR-filtered candidate pool is less inflated and delivers the identification result with higher confidence at the cost of sensitivity.[181]

MODa identified 25 peptides corresponding to 15 proteins for Sample 1 at 1% FDR and 125 peptides corresponding to 51 proteins for Sample 2. Sample 1 contained 5 peptides for AlbM,

while Sample 2 contained 16 peptides (**Table 2.4**). In comparison, MODa and MaxQanut results shared 10 proteins for Sample 1 and 44 proteins for Sample 2 (**Table 2.5** and **Figure 2.5**). However, Samples 1 and 2 only shared 3 identified proteins by both searches: serum albumin (AlbM, P07724), hemoglobin subunit beta-1 (P02088), and heat shock protein HSP 90-beta (P11499).

Notably, the capability of MODa for aligning short sequence tags assigned the candidacy to an AlbM peptide (LPCVEDYLSAILNR) carrying the modification by the biotin dye without FDR-filtering. Corresponding MS/MS spectra of the probe-modified peptide were further verified manually. This modified peptide was observed as $[M + 3H]^{4+}$ ions at 723 *m/z* (**Figure 2.6a**). In-source fragmentation of the attached probe was also observed; a cluster of quadruply charged ions, captured within the 706-732 *m/z* window at the same retention time, shared a significant number of y ions (**Figure 2.7**). ICG molecule is prone to gas-phase fragmentation.[182] However, due to the stochastic sampling of precursor ions for MS/MS in profiling experiments, some intense in-source fragments of the modified peptide were missed for sequencing. Additional forms of the same peptide carrying the intact probe were also observed (**Figure 2.6b to 2.6d**): [singly-oxidized M + $3H]^{4+}$ at 727 *m/z*, $[M + 2H + Na]^{4+}$ at 729 *m/z*, and [doubly-oxidized M + $3H]^{4+}$ at 731 *m/z*. Importantly, all these forms of modified peptide shared a series of y ions (**Figure 2.8**).

Following the same analytical workflow (**Scheme 2.3**), another modified AlbM peptide with a sequence of DTCFSTEGPNLVTR was also discovered (**Figure 2.9** and **Figure 2.10**). Initially, this modified peptide was observed as triply charged ions at *m/z* 927 (**Figure 2.9**). However, the mass increase of this modified peptide matched only a reduced form (loss of the 2-nitro group) of the probe, a known reduction product of 2-nitroimidazole in hypoxic environments.[183] The extraction of ion chromatograms for this peptide, interestingly, also indicated

the presence of several other modified forms of the same peptide (**Figure 2.11**). Their mass increases over the native peptide were all smaller than that of the reduced probe (**Figure 2.10** vs. **Figure 2.9**). Unlike peptide LPCVEDYLSAILNR, those modified forms of peptide DTCFSTEGPNLVTR had different elution times and thus were not in-source fragments.

Mutable modifications of peptide DTCFSTEGPNLVTR were less likely attributed to the degradation of the ICG moiety as part of the reduced probe carried by the peptide. Although ICG was prone to degradation during sample preparations,[176] similar modifications as partial probes were not observed on peptide LPCVEDYLSAILNR. Rather, adducts as partial probes on peptide DTCFSTEGPNLVTR were produced *in vivo*, likely as a result of enzyme-catalyzed degradation of the probe moiety on protein adducts[177] or catabolic processes of the probe in hypoxic tumor preceding or succeeding its conjugation with the protein. Additional ambiguity (**Figure 2.10**) existed for assigning partial probe attachment to the peptide. Mass increases for the peptide overlapped the mass increase for peptide AADKDTCFSTEGPNLVTR, which had a miscleaved site compared to the fully digested peptide DTCFSTEGPNLVTR.

For a few reasons, modification sites for both peptides were assigned to a cysteine residue near the N-terminus, C472 and C591, respectively. First, MS/MS spectra for the modified peptides encompassed y ions up to the C-side residue of cysteine. The sequential y ions denoted both peptides as unique hits against the entire mouse proteome. Second, the cysteinyl thiol (C472) was the only nucleophile in the N-terminal region of peptide LPCVEDYLSAILNR and the most nucleophilic one (C591) in the N-terminal region of peptide DTCFSTEGPNLVTR. Third, no b ion was observed, in agreement with the cysteine modification in the N-terminal region. Additionally, numbers of fragment ions remained unassigned (**Figure 2.8** and **Figure 2.9**), which could be attributed to probe (or probe fragment)-carrying ions generated in the gas phase.[182,184]

Additional experiments can further validate identified peptides and identify additional probe-modified peptides and their precursor proteins. These experiments include targeted proteomics analysis using selected reaction monitoring-tandem mass spectrometry and the all-ion-fragmentation (AIF) for an unbiased measurement of any product ion.[185] DDA spectra and AIF spectra can be explored simultaneously to identify rare protein targets with improved efficiency and confidence. Incorporation of alternative techniques for sample preparation, for instance, on-resin trypsin diegstion[186] and affinity pull-down of probe-modified peptides,[174] is also worth pursuing.

2.3.5 Albumin microenvironment steering pathways for the formation of differential adducts with biotinyl 2-nitroimidazole-ICG

To analyze the probe-protein interaction and visualize the structural significance of peptide LPCVEDYLSAILNR and DTCFSTEGPNLVTR within the folded AlbM, we conducted a computational docking experiment for the biotin dye and AlbM. Although the X-ray crystallographic data of AlbM was not available, the 3D structure was effectively modeled using SWISS-MODEL.[169] SWISS-MODEL is a web-based package for protein structure homology modeling that builds protein models at different levels of complexity. In this study, as the automated mode for modeling was selected, SWISS-MODEL searched for suitable templates against the database of proteins with known 3D structures and identified a list of top hits based on the protein sequence homology. By ranking the identified templates according to the GMQE score, the top hit, rabbit serum albumin (3V09), was selected. The GMQE score described the estimated template-target sequence alignment quality of a specific template and reflected the expected accuracy of its resulting target model and the target's sequence coverage. Afterward, the 3D model of AlbM was downloaded as a PDB file and appended hydrogen atoms. Meanwhile, all chemical

bonds in the biotin dye model except those forming ring and planar structures were defined as flexible bonds. Finally, the prepared AlbM and biotin dye models were used together with a defined search space at either the center or the peripheral subdomain of the protein model as the docking input for AutoDock Vina.[171] Subsequently, AutoDock Vina computed the free energy of various ligand-bound protein models and searched for the best docking mode with its relative free energy at the local minimum. By placing the ligand at various locations on the protein, an affinity value regarding binding energy was calculated for each docked model. AutoDock Vina reported nine docked modes with the ligand affinity ranging from -8.0 to -8.5 kcal/mol for peptide LPCVEDYLSAILNR and -5.4 to -6.1 kcal/mol for peptide DTCFSTEGPNLVTR. Such energy differences were equivalent to 20- to 100-fold differences in the dissociation constant (K_d). The structure and location profiles of the docked ligand were saved as the output (**Figure 2.12**).

Intriguingly, locations of peptide LPCVEDYLSAILNR containing C472 and DTCFSTEGPNLVTR containing C591 were significantly distinct on the 3D model of AlbM. Peptide LPCVEDYLSAILNR (with C472 in red, **Figure 2.12**) was located in a deep pocket, where the biotin dye was docked with a higher binding affinity. In contrast, peptide DTCFSTEGPNLVTR (with C591 in yellow, **Figure 2.12**) was located on the surface of the protein, where the biotin dye was docked with a lower binding affinity. The homology analysis (**Table 2.6**) of these sequence regions validated their relevant template-deduced conformational features of the AlbM model. Hence, different locations of the two cysteine residues could explain the differential formation of probe adducts: the intact biotin dye molecule modified the inner peptide while a reduced probe or probe fragments modified the surface peptide (**Scheme 2.4**).

Upon intravenous administration, the biotin dye could exist in two hypothetical forms as complexed with the high-abundance AlbM in the circulatory system of the mouse subject: (1)

AlbM peripheral subdomain-bound form and (2) AlbM core pocket-buried form.[187,188] The AlbM might carry the probe in both forms through the mouse circulatory system to the tumor site. The high affinity of the probe to the pocket increases the effective molarity[189] of the probe, thus amplifying its apparent electrophilicity, i.e., making the otherwise inert 2-nitroimidazole ring reactive towards the reduced cysteine residue (C472) that is made available by the compromised redox homeostasis[190] at the hypoxic tumor site.

On the other hand, the peripheral subdomain-interacting or surface-adsorbed 2-nitroimidazole probe was exposed to the solvent environment, prone to dynamic dissociation and re-association, accessible to oxidoreductases for enzymatic reduction. When nucleophilic thiol groups become available from impaired disulfide bridges of AlbM at the hypoxic tumor site, the thiol group on cysteine (C591) reacts with the activated imidazole ring on the reduced probe and metabolic fragments via the nucleophilic addition or substitution mechanism,[156,188] leading to observed modifications of different mass increases on peptide DTCFSTEGPNLVTR (**Figure 2.9** and **Figure 2.10**). Modifications of this peptide exemplify a general pathway for 2-nitroimidazole modification of proteins in the hypoxic tumor (**Scheme 2.4**).

Besides the hypoxic condition, the biodistribution of albumin could play another significant role in the fluorescence detection of solid tumors in this study. While 49% of the total albumin population is located within the plasma as intravascular albumin or “serum” albumin after being synthesized and excreted by hepatocytes, a substantial 51% of it is found as extravascular albumin at discrete levels of local abundance.[191] Importantly, the tumor cell line 4T1-Luc used in this study is a breast cancer cell line, which originates from malignant neoplasms of the mouse mammary gland. With lactation as its unique function, the mammary gland can not only capture albumin from the plasma for milk secretion via transcytosis[192] but also synthesize nonhepatic

albumin on its own[193]. In comparison to its scare intracellular abundance of most somatic cells, albumin occupies 19% of cytosolic protein contents in breast cancer cells.[194] In fact, the level of this intracellular albumin has been utilized as a prognostic factor to evaluate the effect of adjuvant tamoxifen (whose active metabolites can block estrogen receptors that sustain the growth of breast cancer cells) treatment for ER+ (estrogen receptor-positive) breast cancer.[195] Therefore, it is conceivable that the malignant breast tumor behaves as a "hypoxic reservoir" of albumin, which provides an exclusive extra- and intracellular environment where the low abundance of oxygen and high abundance of albumin coexist. It is this unique pathophysiological environment that enables the hypoxia-targeting capability of 2-nitro-ICG probes.

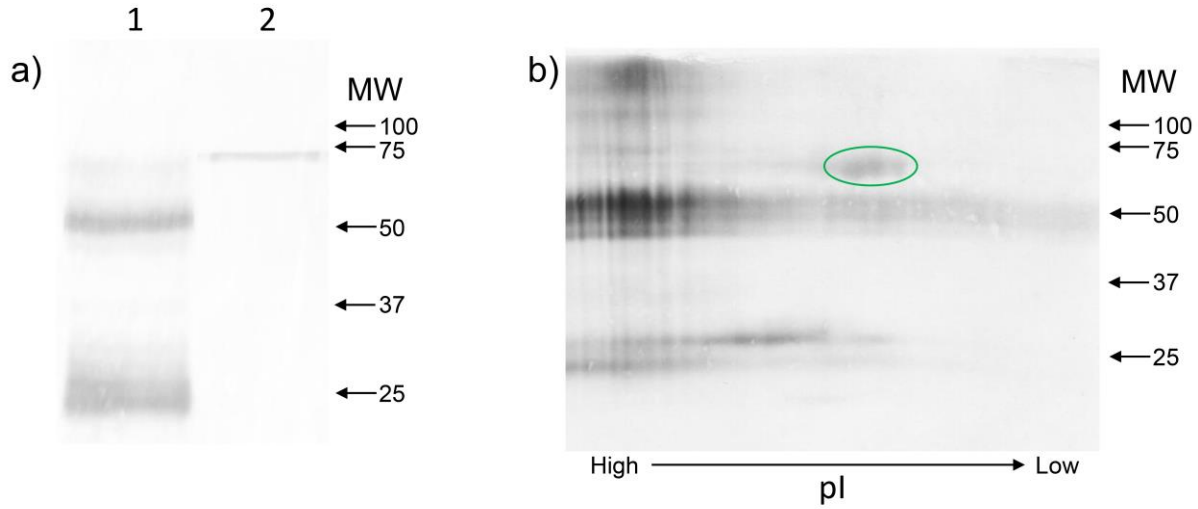
The probe molecule would likely to be transiently engaged by a protein carrier, such as AlbM, transported through the cardiovascular system to various destinations, activated locally by an oxidoreductase, and eventually bond to the protein carrier forming covalent adducts under hypoxic condition. Overall, the discovery of two biotin dye-modified AlbM peptides did bestow some valuable insights on the hypoxia-detection mechanism of 2-nitroimidazole derivatives (**Scheme 2.4**).

2.4 Conclusion

With the biotinyl 2-nitroimidazole-ICG probe, we have opened a proteomics gate to explore the metabolic chemistry of 2-nitroimidazole and its hypoxia-targeting mechanism in the hypoxic cellular condition. By applying our product-ion-oriented data-mining methodology, we have successfully interpreted otherwise convoluted and indecisive data of proteomics profiling. Accordingly, the biotinyl 2-nitroimidazole-ICG probe and its reduced forms modify the mouse serum albumin as the primary protein target at two cysteine residues, C472 and C591, located in a deep pocket and on the surface, respectively. At the hypoxic tumor site of the probe-injected mouse subject, the chemical probe covalently links to these peptides probably via dissimilar mechanisms (**Scheme 2.4**). On top of this constructive interpretation to the global insight on the hypoxia-probing mechanism of 2-nitroimidazole compounds, our investigation illustrates a distinguished analytical workflow that depicts a practical guideline for some foreseeable challenging cases in the mass spectrometry-based chemical proteomics, especially during *in vivo* evaluation of drug candidates, where the detectable chemical modifications on target and off-target proteins are unknown or mutable.

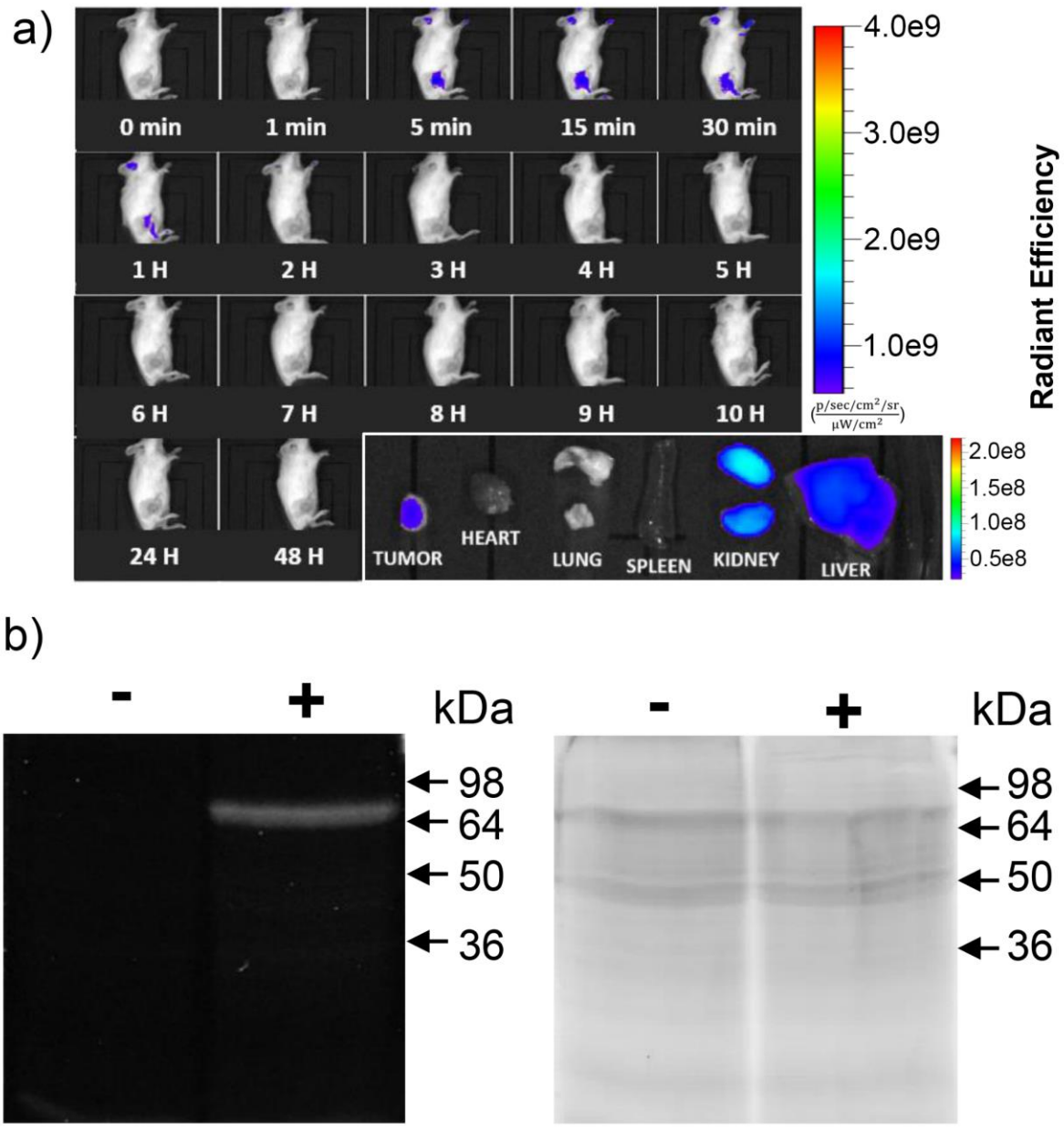
2.5 Chapter 2 Figures

Figure 2.1 Western Blot images.



Note: a) WB result for SDS-PAGE gel of unenriched probe-treated sample as Lane 1 and immobilized avidin-enriched probe-treated sample as Lane 2. **b)** WB result for 2-DE gel of unenriched probe-treated sample. The protein spot of interest is marked in the green ellipse. The WB experiments involve anti-biotin primary antibody and alkaline phosphatase conjugate as the secondary antibody for chromogenic detection.

Figure 2.2 Fluorescence imaging results.



Note: a) *in vivo* (top), *ex vivo* (bottom) fluorescence images of tumor sample labeled by the biotin dye. **b)** Fluorescence (left) and Coomassie Blue G250-stained (right) gel images of SDS-PAGE-resolved tumor lysate samples (lane "-" is control and lane "+" is biotin dye-labeled tumor sample). Refer to **Figure 2.10** for additional gel images.

Figure 2.3 Quantitative comparison of in vivo tumor fluorescence peak values of different groups of mice injected with different dyes.

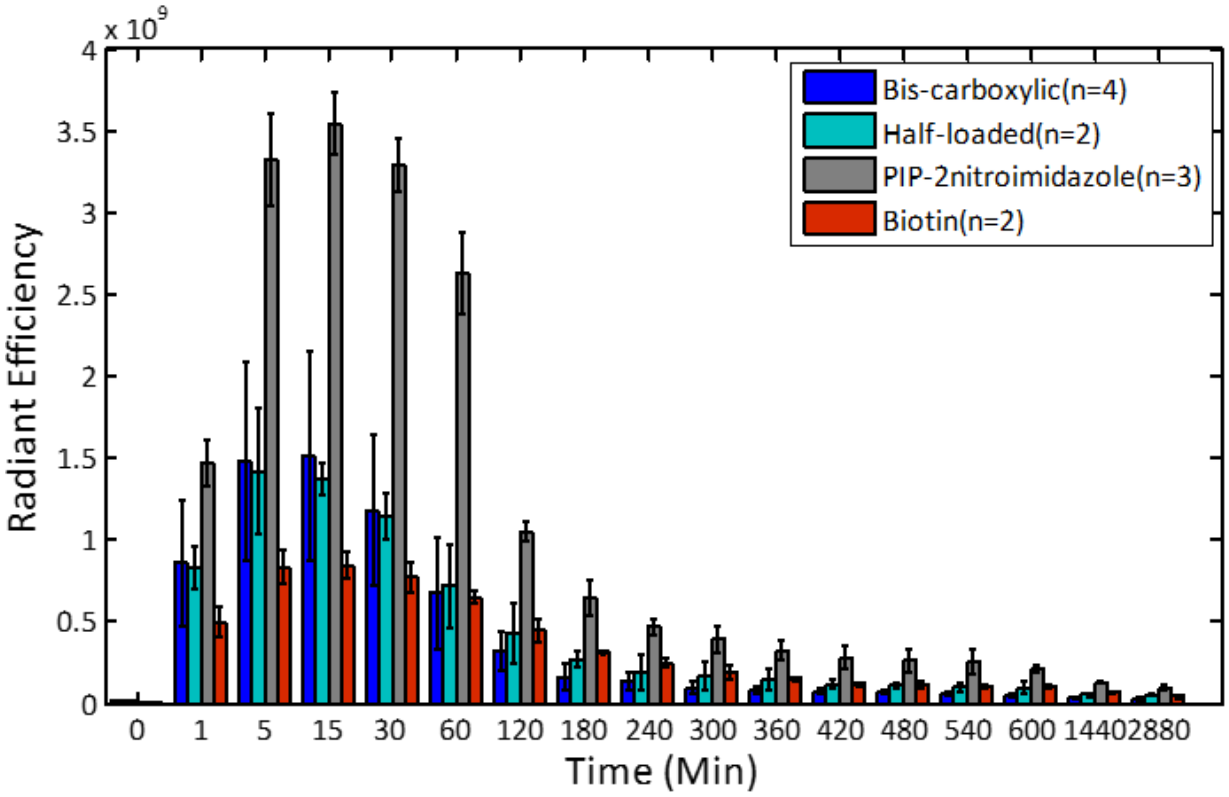
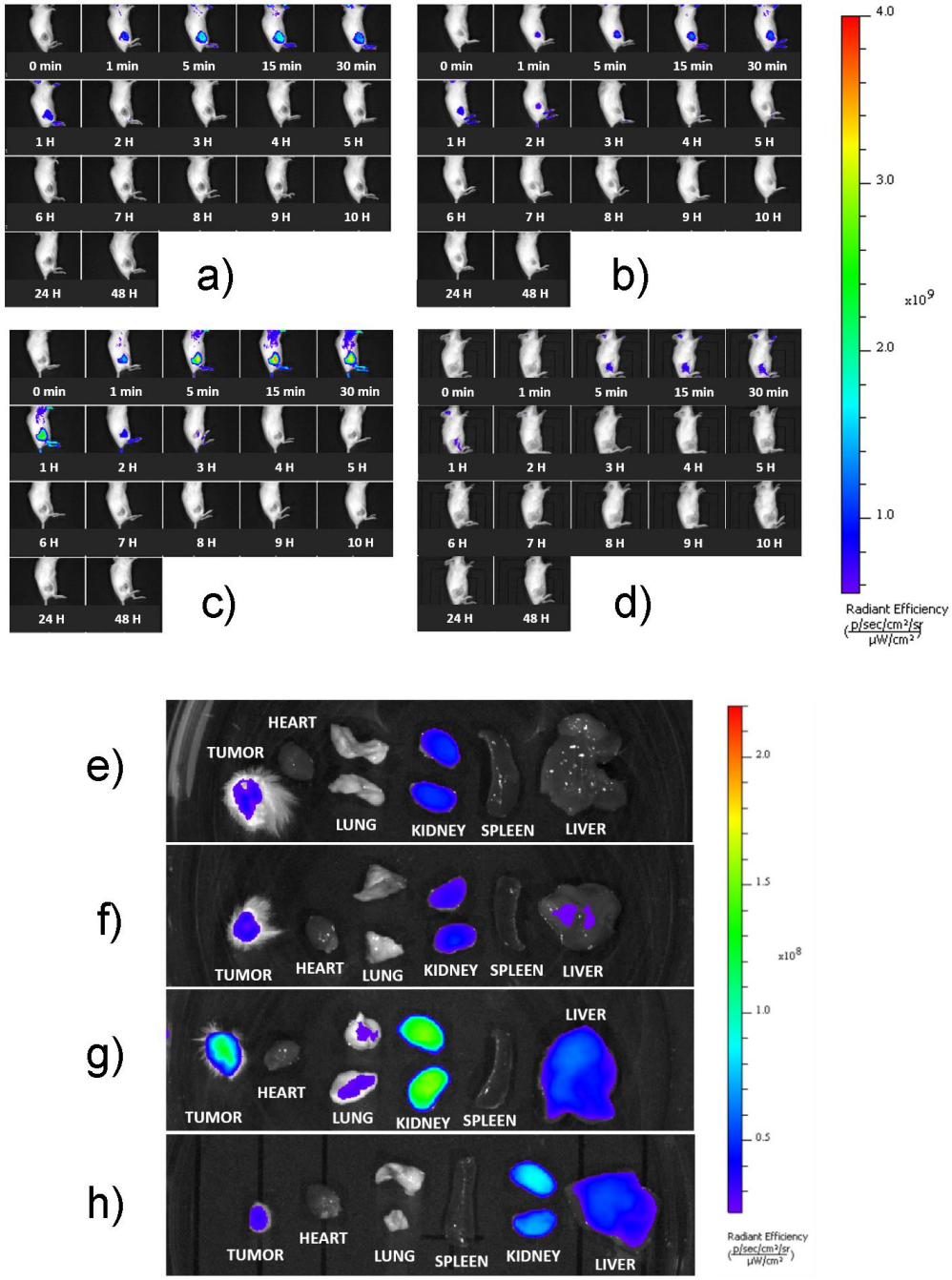
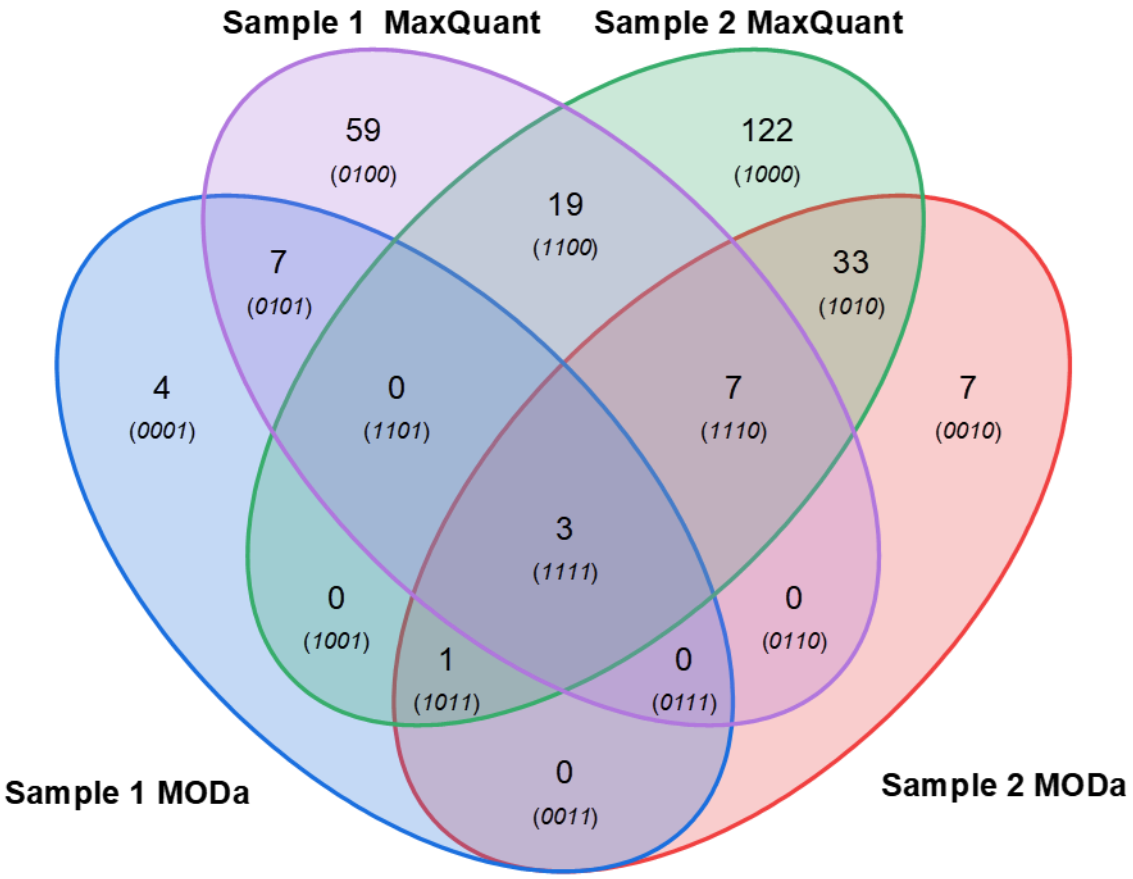


Figure 2.4 *in vivo* and *ex vivo* Fluorescence kinetics of individual mice before dye injection and at different time points after dye injection.



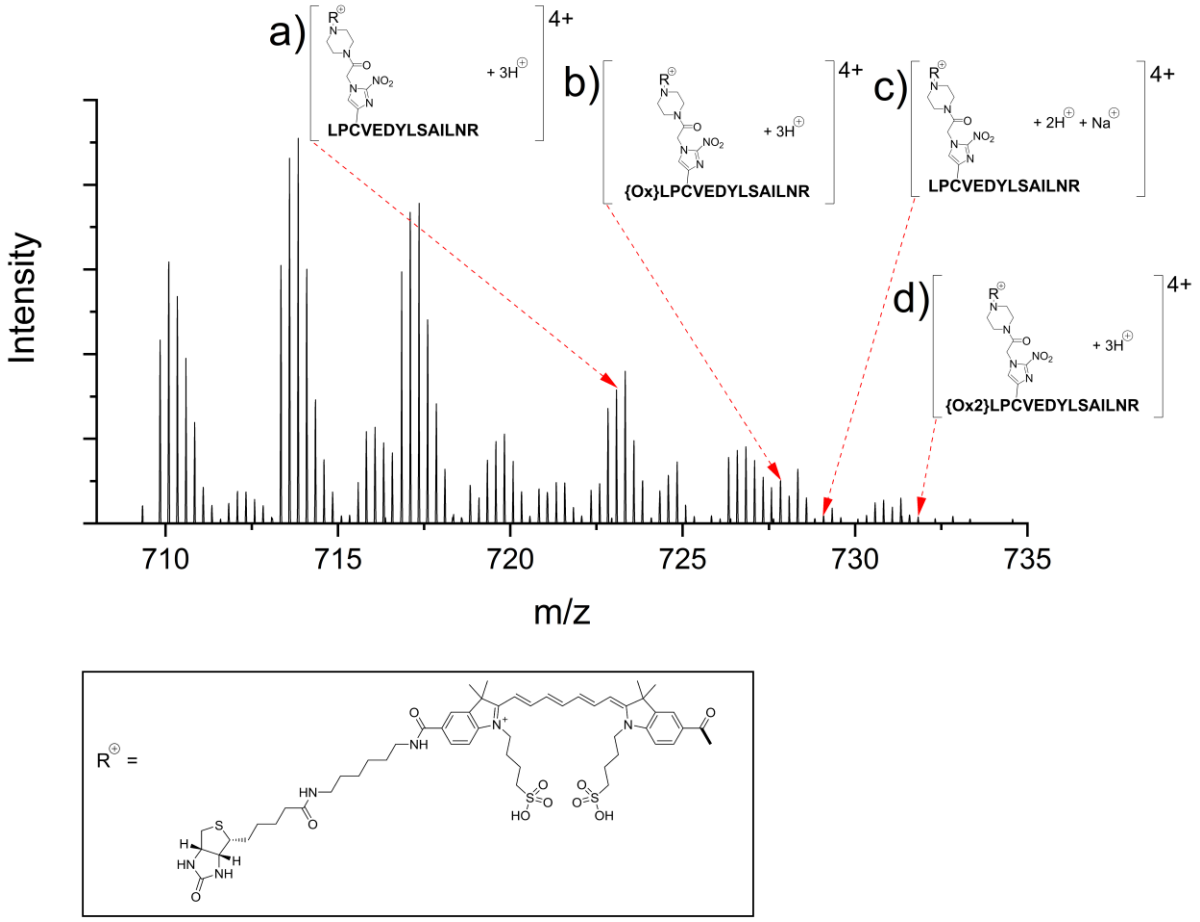
Note: Each group of mice is injected with one of four dyes: **(a)** and **(e)** Bis-carboxylic ICG, **(b)** and **(f)** Half-loaded ICG, **(c)** and **(g)** Pip-2nitroimidazole-ICG and **(d)** and **(h)** Biotin dye, with 100 µl at 25 µM concentration solved in 9.25% sucrose solution. For each mouse, the tumor is located on top of the right leg of the mouse.

Figure 2.5 Venn diagram showing relations among the identification results.



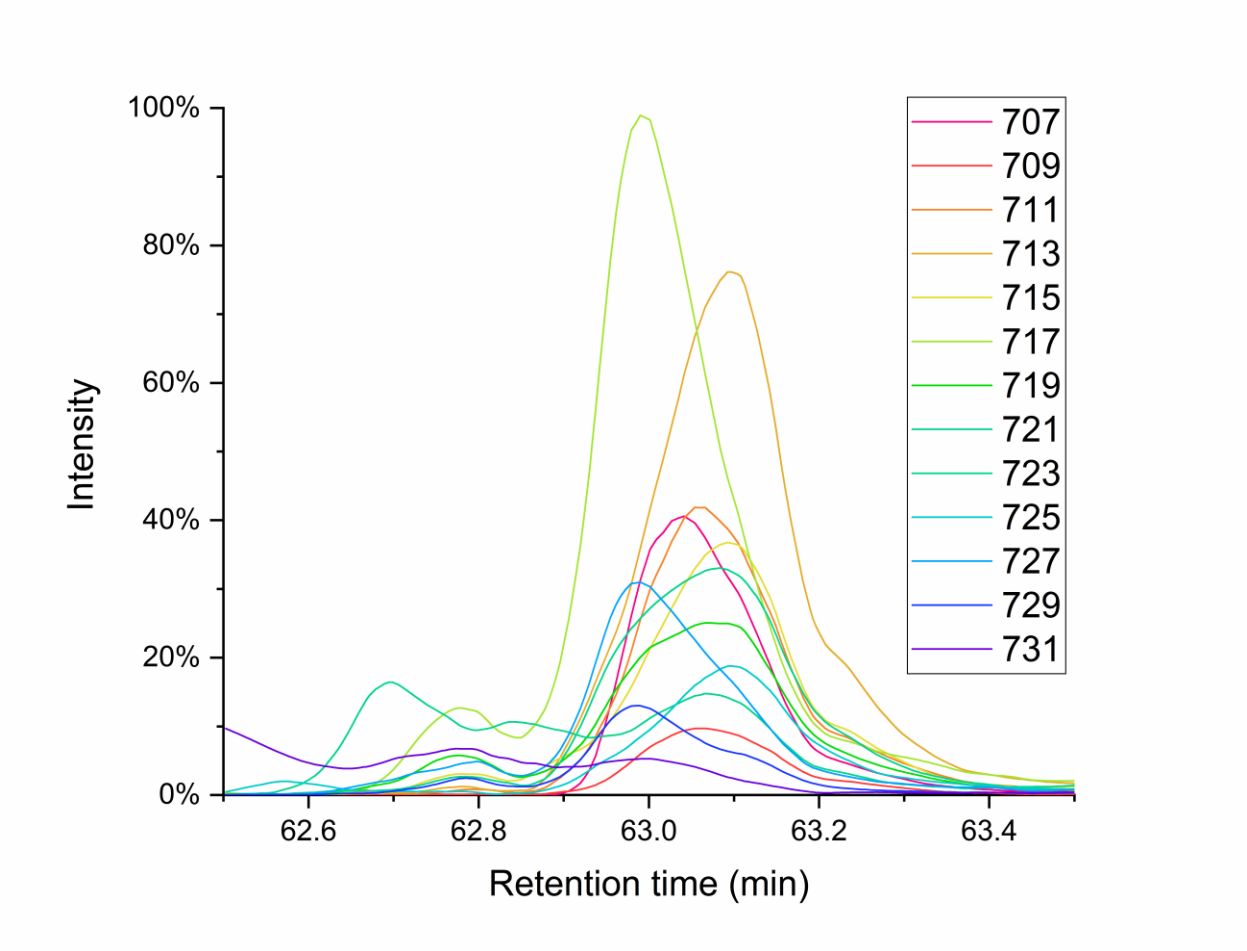
Note: This Venn diagram shows relations among the identification results of both SDS-PAGE-resolved avidin-enriched probe-treated sample (Sample 1) and 2-DE-resolved unenriched probe-treated sample (Sample 2) by both MODa and MaxQuant. Refer to **Table 2.5** for specific protein identities labeled with 4-digit binary numbers.

Figure 2.6 Example MS spectrum associated with the peptide LPCVEDYLSAILNR modified by intact biotin dye.



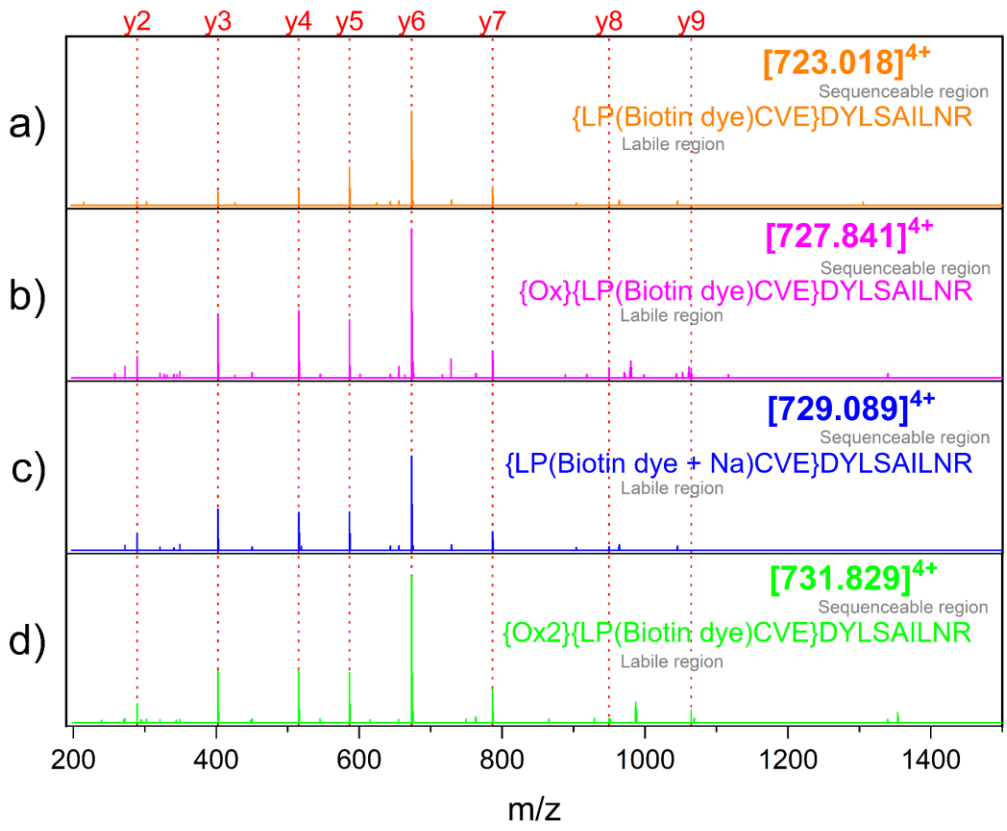
Note: This example MS spectrum shows a cluster of quadruply-charged molecular ions (labeled) and their in-source fragmented forms (unlabeled) associated with the peptide LPCVEDYLSAILNR modified by intact biotin dye. **a)** The peak at *m/z* of 723 matching the probe-carrying peptide. **b)** The peak at *m/z* of 727 matching the oxidized (plus one oxygen atom) form of a probe-carrying peptide. **c)** The peak at *m/z* of 729 matching the one sodium adduct form (replacing one proton with one sodium) of a probe-carrying peptide. **d)** The peak at *m/z* of 731 matching the doubly oxidized (plus two oxygen atoms) form of a probe-carrying peptide.

Figure 2.7 Alignment of extracted ion chromatograms for the MS ion cluster associated with peptide LPCVEDYLSAILNR modified by the intact probe.



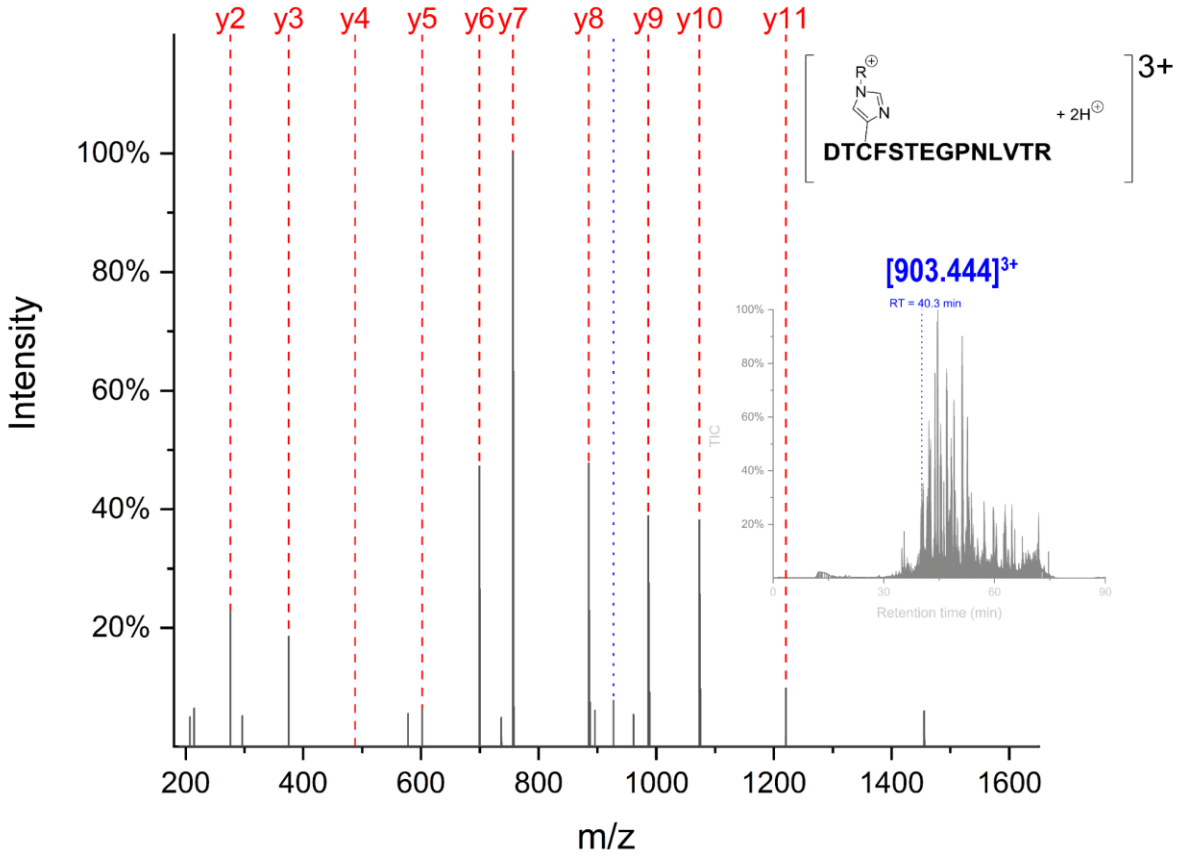
Note: The 706-732 *m/z* ion isolation window of interest is binned into smaller windows of 2 *m/z*. The overlap of ion current peaks implies multiple ions share the same retention profile, suggesting the lability of the probe modification and in-source fragmentation phenomenon of the original probe-carrying peptide.

Figure 2.8 Annotated MS/MS spectra featuring the peptide LPCVEDYLSAILNR modified by intact biotin dye.



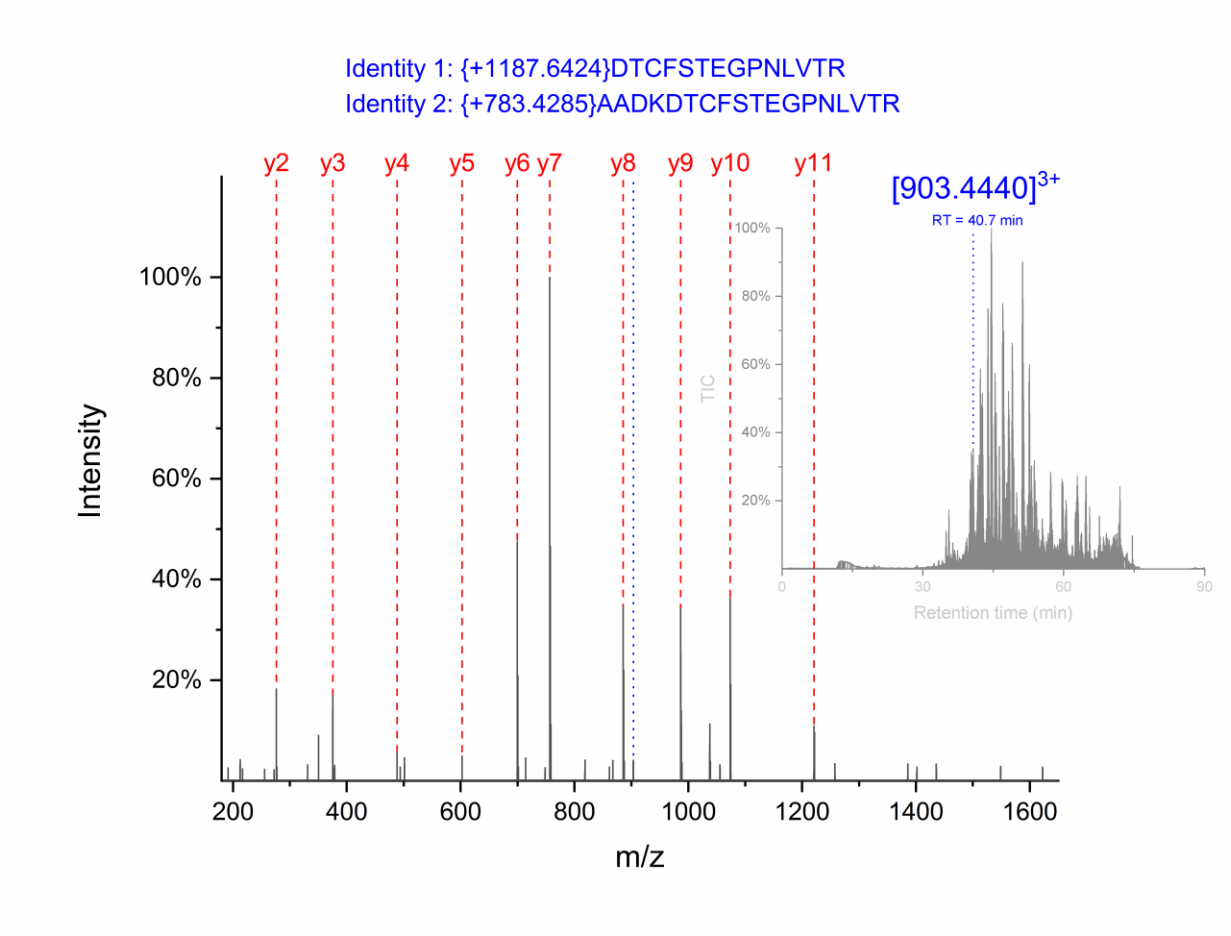
Note: a) The spectrum matching the probe-carrying peptide. **b)** The spectrum matching the oxidized (plus one oxygen atom) form of the probe-carrying peptide. **c)** The spectrum matching the one sodium adduct form (replacing one proton with one sodium) of the probe-carrying peptide. **d)** The spectrum matching the doubly oxidized (plus two oxygen atoms) form of the probe-carrying peptide. Oxidation of probe-carrying peptides **b** and **d** was peculated to occur during sample preparation.

Figure 2.9 Annotated MS/MS spectrum featuring the peptide DTCFSTEGPNLVTR modified by a reduced form of the biotin dye.



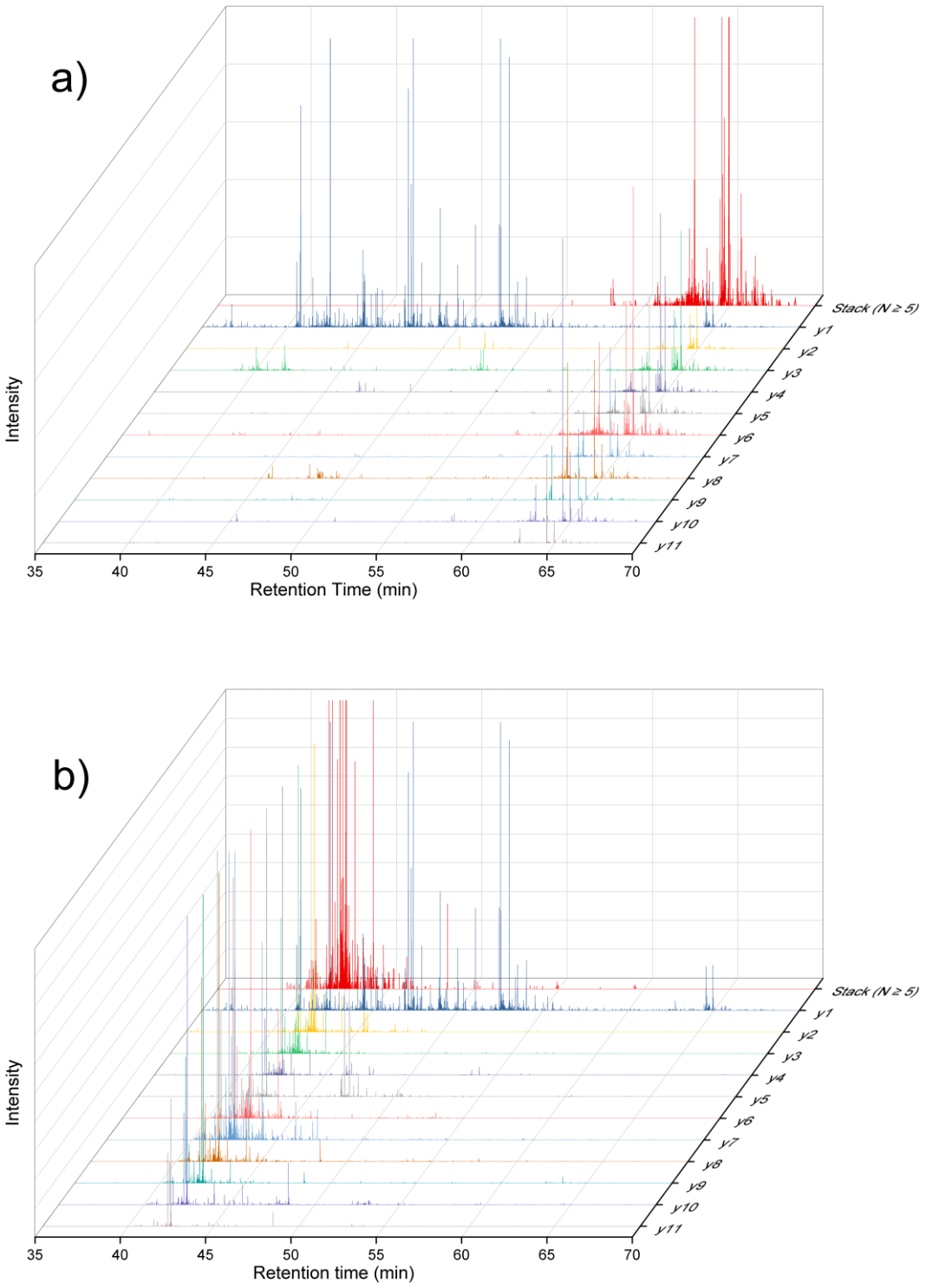
Note: This triply charged precursor with an *m/z* ratio of 927 matches a proposed modified peptide carrying a reduced probe with a loss of 2-nitro group on its imidazole ring. This peptide eluted at 40.3 min as shown on the total ion chromatogram (TIC). Refer to **Scheme 2.4** for the complete structural scheme of the modification (fully protonated, R^+).

Figure 2.10 Annotated MS/MS spectrum featuring the probe-modified peptide DTCFSTEGPNLVTR or its mis-cleaved form.



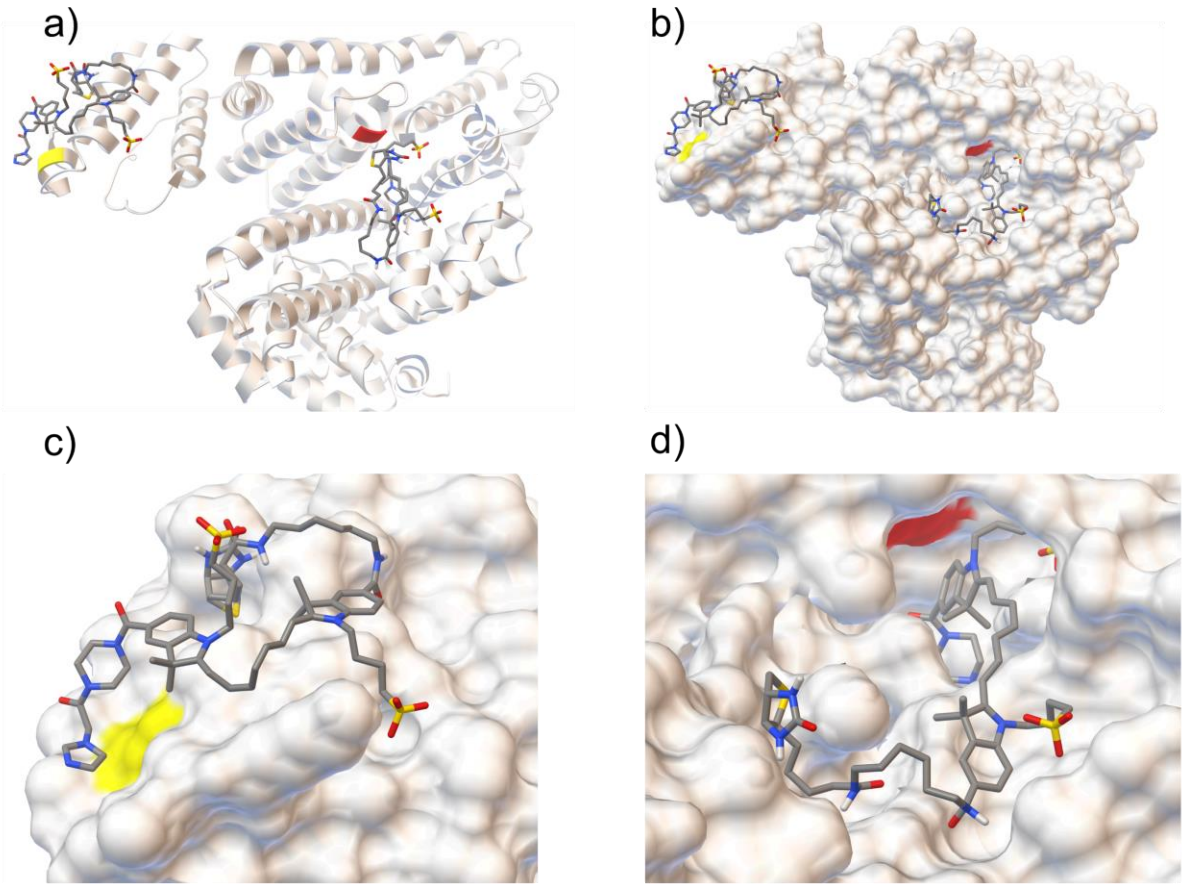
Note: This annotated MS/MS spectrum features either the probe-modified peptide DTCFSTEGPNLVTR or its mis-cleaved form AADKDTCFSTEGPNLVTR. This triply charged precursor has a *m/z* ratio of 903. This precursor eluted at 40.7 min as shown on the total ion chromatogram (TIC).

Figure 2.11 Alignment of extracted ion chromatograms (XICs) for y ions of interest.



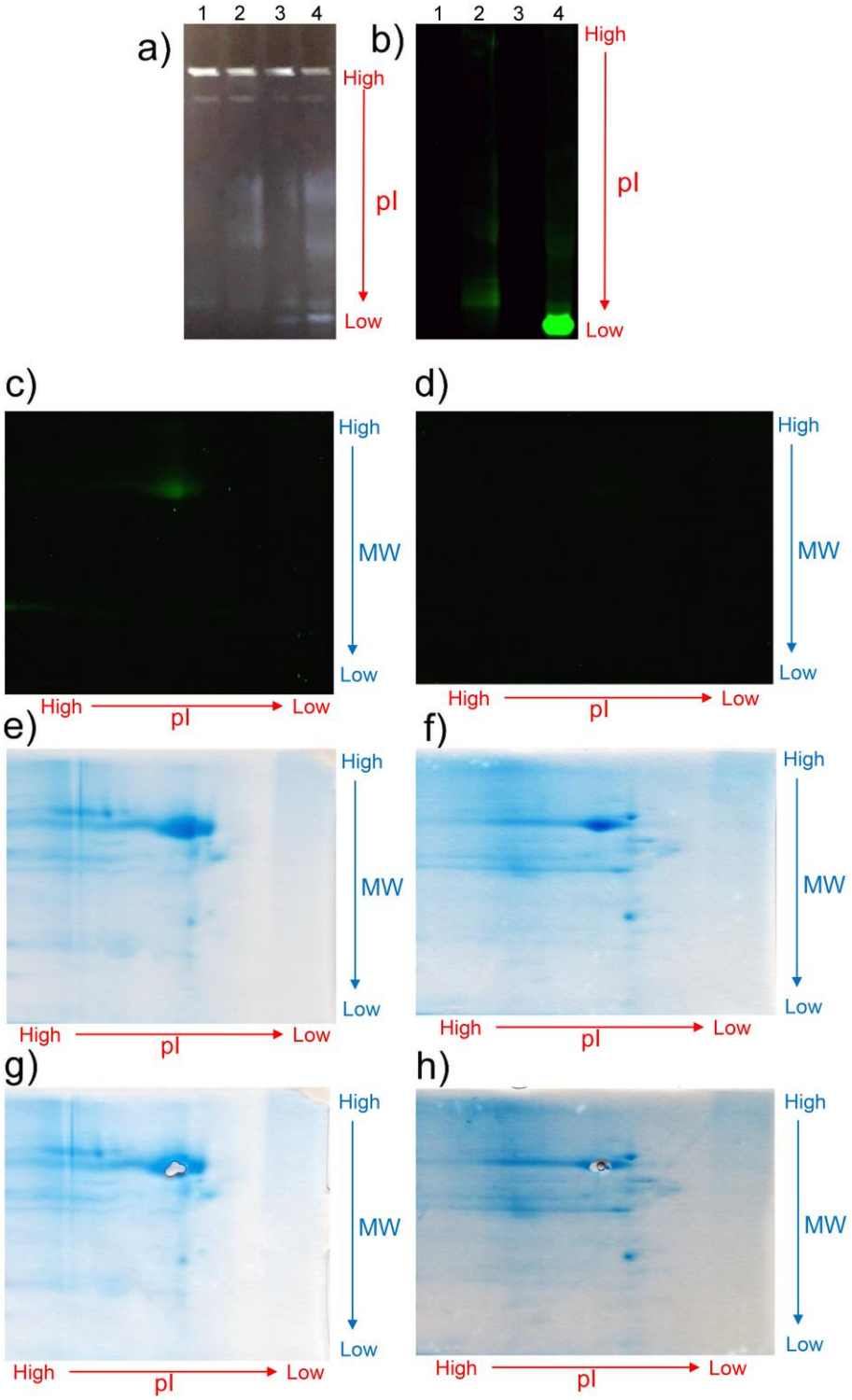
Note: a) Aligned y-ion XICs of peptide LPCVEDYLSAILNR. **b)** Aligned y-ion XICs of (AADK)DTCFSTEGPNLVTR. The alignment shows both individual (y1 to y11) and stacked (showing as summed y-ion current for MS/MS scans with at least five concurrent y ions) ion chromatograms. Y-axis was zoomed for a better illustration with a scale of 1 x 10^7. On the stack chromatograms, one peak cluster indicates either a single molecular ion or multiple molecular ions with similar precursor masses (exemplifying in-source fragmentation) occurring at a specific retention window at 63 min, referring to **Figure 2.7**. The dispersion of multiple peak clusters over the retention time (43 min, 48 min, 50 min, 55 min, and 59 min on **b** suggests the existence of multiple molecular ions associated with this peptide of interest.

Figure 2.12 Biotin dye docking results visualizing the pre-modification probe-protein interactions.



Note: The structure of AlbM was modeled using a template X-ray structure (rabbit serum albumin PDB accession 3V09). **a)** Cartoon presentation of AlbM binding the biotin dye molecule in two distinct manners. The cysteine residue is highlighted in yellow for the superficial peptide DTCFSTEGPNLVTR and red for the conserved peptide LPCVEDYLSAILNR. **b)** Protein-surface presentation with docked biotin dye molecules at both the surface and cavity of AlbM. **c)** A regional view of the biotin dye-protein interaction on the binding surface. **d)** A regional view of the biotin dye-protein interaction in the binding pocket. The 2-nitroimidazole group is concealed inside the pocket and isolated from the extra-molecular environment.

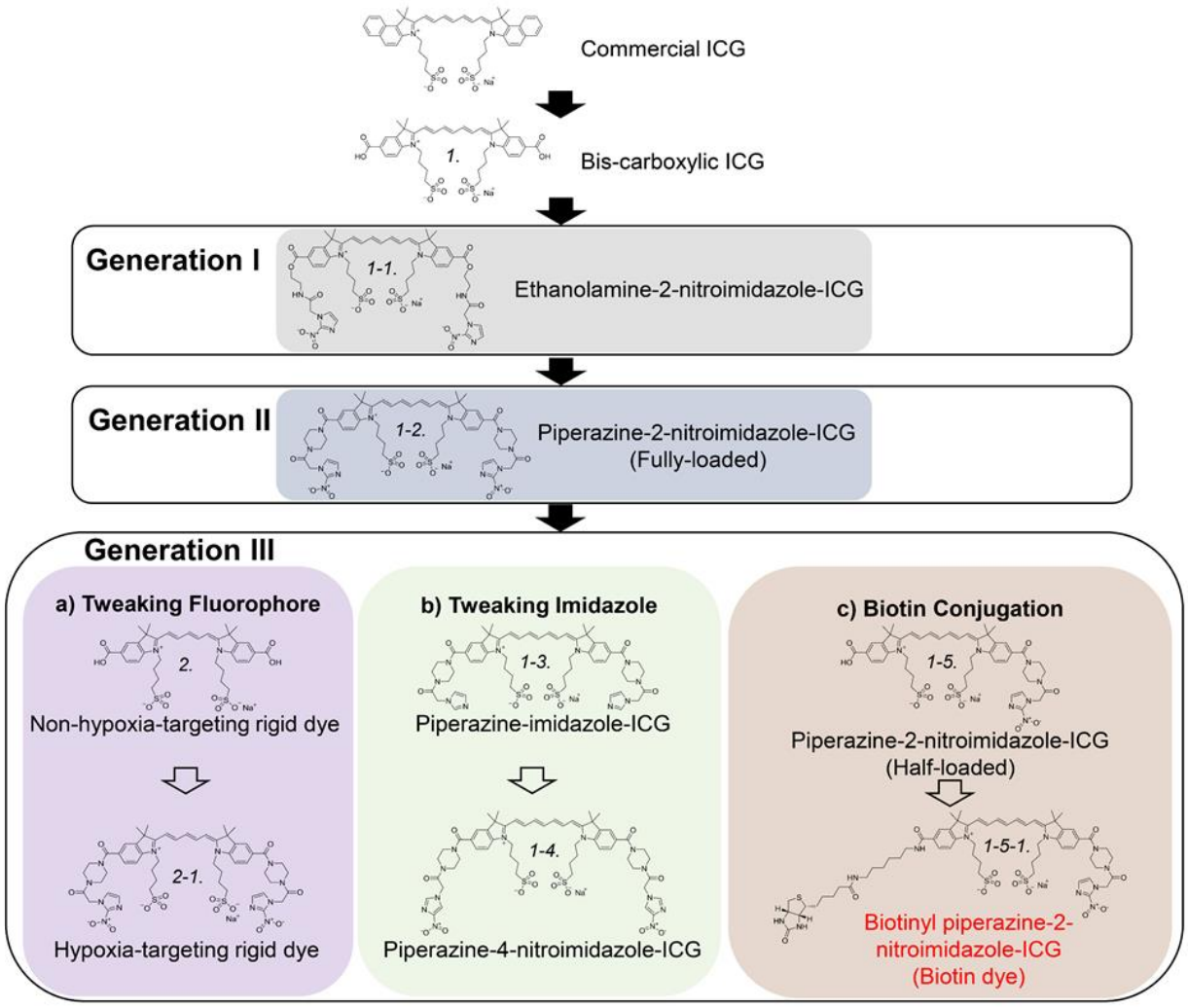
Figure 2.13 Additional gel images.



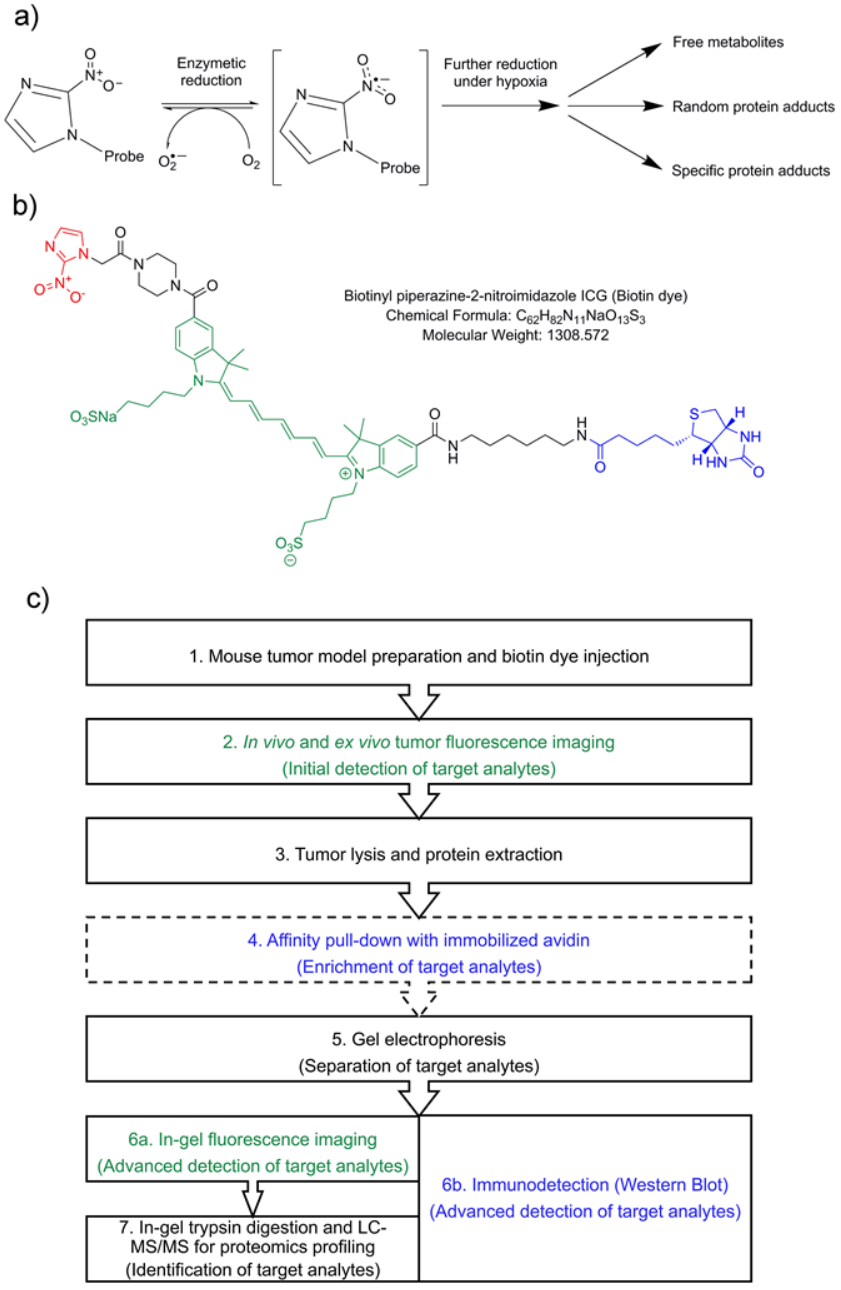
Note: a) Visible bands in vIEF-PAGE gel after TCA-induced protein precipitation. Lane 1: denatured control sample, Lane 2: denatured probe-treated sample, Lane 3: native control sample, and Lane 4: native probe-treated sample. **b)** Fluorescence image of **a**. **c)** 2-DE fluorescence image of a denatured probe-treated sample. **d)** 2-DE fluorescence image of **a** denatured control sample. **e)** Coomassie Blue-stained 2-DE gel of denatured probe-treated sample. **f)** Coomassie Blue-stained 2-DE gel of denatured control sample. **g)** Gel **e** post-excision of the protein spot of interest. **h)** Gel **f** post-excision of the protein spot of interest.

2.6 Chapter 2 Schemes

Scheme 2.1 Evolution of nitroimidazole-indocyanine green derivatives as fluorescent tumor hypoxia probes.

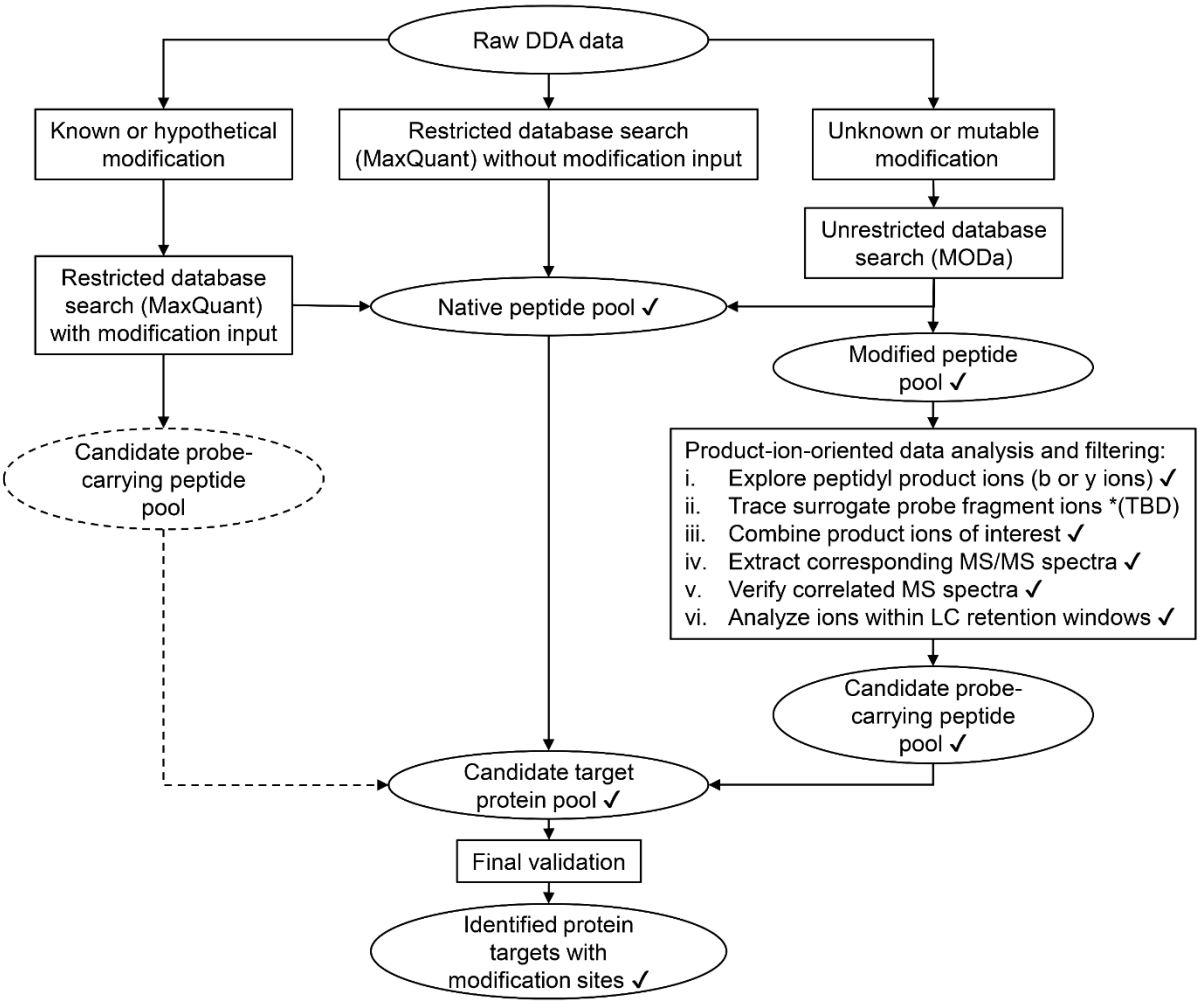


Scheme 2.2 Overview of this investigation.



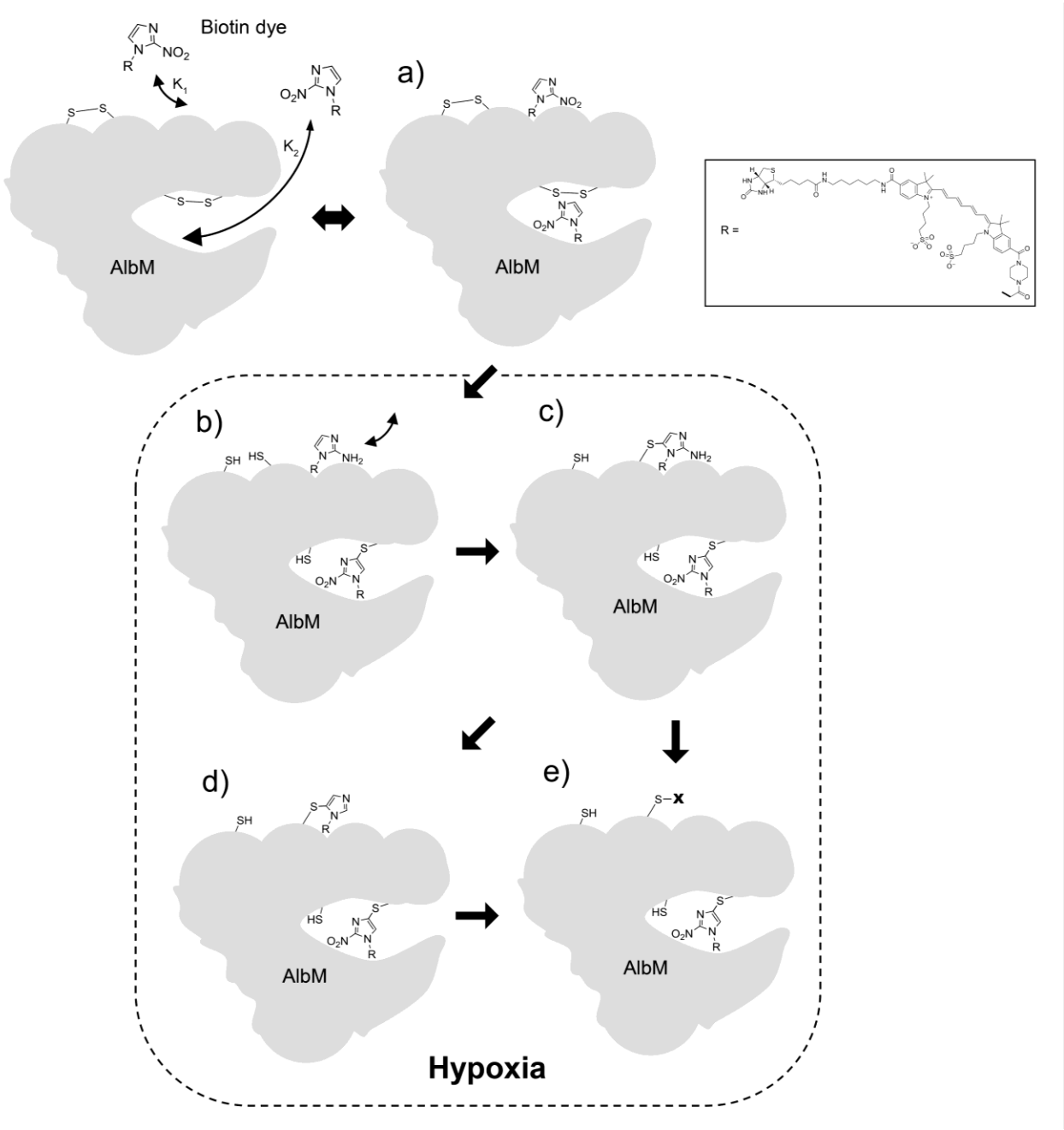
Note: a) Brief scheme showing proposed general reaction pathways of 2-nitroimidazole probes. Both small-molecule metabolites and macromolecule adducts are anticipated as in vivo products. **b)** Structure of biotin dye showing the reactive hypoxia-targeting 2-nitroimidazole group in red, ICG fluorophore in green, and biotin affinity handle in blue. **c)** Experimental workflow of this study. Note the biotin dye is dual-functional thus allows for flexible and orthogonal detection of target analytes at various stages (color-coded to indicate related groups on the dye molecule).

Scheme 2.3 Recommended data-mining workflow for the identification of protein targets and their modification sites in chemical proteomics profiling studies.



Note: The mass spectra acquired in data-dependent acquisition mode should be analyzed with complementary tactics in parallel, regarding the identifiability of modified peptides. Using this study as an example, the output corresponding to each step of data processing has been verified (✓). The dashed arrow represents the prevalent restrictive search strategy relying on an input of defined modifications to assign spectra featuring modified peptides, which has been proven inapplicable to this study. Traceable surrogate probe fragment ions remain to be determined *(TBD), which requires a dedicated MS/MS study of the biotin dye in the future.

Scheme 2.4 Proposed AlbM-biotin dye *in vivo* reaction mechanism.

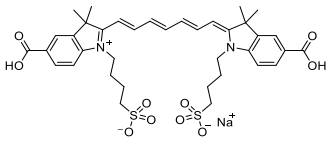
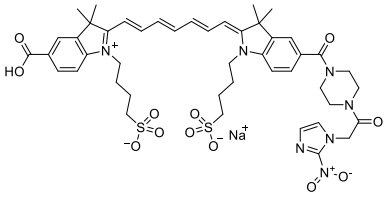
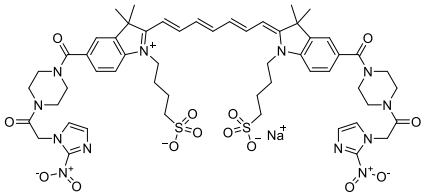
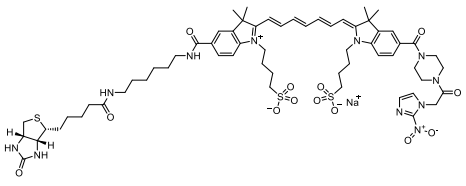


Note: The AlbM carries the biotin dye in two hypothetical forms (regardless of stoichiometry), the peripheral subdomain-bound (surface-adsorbed) and core pocket-buried, to the hypoxic tumor site through the circulatory system of the mouse subject. These two forms of the protein-probe complex theoretically have a 20- to 100-fold difference in equilibrium constants for the probe binding (K_1 vs. K_2). Under hypoxic condition, as the compromised redox homeostasis impairs proximate disulfide bridges of AlbM cysteine residues, both forms of the probe can covalently modify AlbM. Note the high affinity of the probe to the protein pocket elevates the effective molarity of the probe thus amplifying its apparent electrophilicity,

which provides additional reactivity of the probe to the adjacent thiol in the binding pocket, leading to direct addition of the intact probe to the cysteine residue (center **a** and **b**). In contrast, the surface-adsorbed probe is only reactive to the adjacent thiol after the bio-reductive activation (top, **b**). Furthermore, post-covalent-binding bio-reductive reactions reduce 2-aminoimidazole (top, **c**) to imidazole (top, **d**) or other unknown species (X, top, **e**) at the protein surface.

2.7 Chapter 2 Tables

Table 2.1 Optical properties of the biotin dye in comparison with the previous ICG-based fluorescence imaging probes.

Name	Structure	Chemical Formula	$\lambda_{absorption\ max}$ (nm)	$\lambda_{emission\ max}$ (nm)	Extinction Coefficient (x1000 $M^{-1}cm^{-1}$)
Bis-carboxylic ICG		$C_{37}H_{43}N_2NaO_{10}S_2$	755	790	221
Half-loaded ICG		$C_{46}H_{54}N_7NaO_{12}S_2$	755	790	171
Pip-2-nitroimidazole-ICG (Fully loaded)		$C_{55}H_{65}N_{12}NaO_{14}S_2$	753	790	230
Biotin dye		$C_{62}H_{82}N_{11}NaO_{13}S_3$	755	785	102

Note: Optical properties of the biotin dye are compared with the previous ICG-based fluorescence imaging probes.[148,150] For the measurement of optical properties, 9.25% of sucrose was used as the solvent in consistence with published results. The fluorescence signal was more stable in this solvent. The fluorescence signal of ICG dye varies greatly in different solvents. See **2** for more detailed references.

Table 2.2 AlbM peptides identified by MaxQuant.

Identified Sample1 (SDS-PAGE) Peptide Sequence	Identified Sample2 (2-DE) Peptide Sequence
AADKDTCFSTEGPNLVTR	AADKDTCFSTEGPNLVTR
AETFTFHSDICTLPEK	AETFTFHSDICTLPEK
AETFTFHSDICTLPEKEK	AETFTFHSDICTLPEKEK
AHCLSEVEHDTMPADLPAIAADFVEDQEVCK	AHCLSEVEHDTMPADLPAIAADFVEDQEVCK
ALVSSVR	ALVSSVR
APQVSTPTLVEAAR	APQVSTPTLVEAAR
AWAVAR	AWAVAR
CCAEANPPACYGTVLAEFQPLVEEPK	CCAEANPPACYGTVLAEFQPLVEEPK
CCSGSLVER	CCSGSLVER
CCTLPEDQRLPCVEDYLSAILNR	CCTLPEDQRLPCVEDYLSAILNR
DVFLGTFLYEYSR	DVFLGTFLYEYSR
ECCHGDLLECADDR	ECCHGDLLECADDR
EFKAETFTFHSDICTLPEKEK	EFKAETFTFHSDICTLPEKEK
ENPTTFMGHYLHEVAR	ENPTTFMGHYLHEVAR
ENYGELADCCTK	ENYGELADCCTK
GLVLIAFSQYLQK	GLVLIAFSQYLQK
HPDYSVSLLLR	HPDYSVSLLLR
KQTALAELVK	KQTALAELVK
LCAIPNLR	LCAIPNLR
LCAIPNLRENYGELADCCTK	LCAIPNLRENYGELADCCTK
LGEYGFQNAILVR	LGEYGFQNAILVR
LPCVEDYLSAILNR	LPCVEDYLSAILNR

LQTCCDKPLLK	LQTCCDKPLLK
RHPDYSVSLLLR	RHPDYSVSLLLR
RPCFSALTVDETYVPK	RPCFSALTVDETYVPK
RPCFSALTVDETYVPKEFK	RPCFSALTVDETYVPKEFK
SLHTLFGDK	SLHTLFGDK
SLHTLFGDKLCAIPNLR	SLHTLFGDKLCAIPNLR
SLHTLFGDKLCAIPNLRENYGELADCCTK	SLHTLFGDKLCAIPNLRENYGELADCCTK
TVMDDFAQFLDTCCK	TVMDDFAQFLDTCCK
VCLLHEK	VCLLHEK
VNKECCHGDLLECADDRAELAK	VNKECCHGDLLECADDRAELAK
YNDLGEQHFK	YNDLGEQHFK
YTQKAPQVSTPTLVEAAR	YTQKAPQVSTPTLVEAAR
ATAEQLK	AADKDTCFSTEGPNLVTRCK
CCTLPEDQR	AFKAWAVAR
CSYDEHAK	ATAEQLKTVMDDFAQFLDTCCK
EAHKSEIAHR	DDNPSLPPFERPEAEAMCTSFK
KAHCLSEVEHDTMPADLPAIAADFVEDQEVCK	DTCFSTEGPNLVTR
KYEATLEK	ECCHGDLLECADDRAELAK
LATDLTK	EFKAETFTFHSDICTLPEK
LQTCCDKPLLKK	EKALVSSVR
QTALAELVK	LCAIPNLRENYGELADCCTKQEPER
SEIAHR	LDGVKEKALVSSVR
TCVADESAANCDK	LSQTFPNADFAEITK
TNCDLYEK	LSQTFPNADFAEITKLATDLTK
YMCENQATISSK	LVQEVTDFAK

	NECFLQHKDDNPSLPPFERPEAEAMCTSFK
	NYAEAKDVFLGTFLYEYSR
	QEPERNECFLQHKDDNPSLPPFERPEAEAMCTSFK
	QTALAELVKHKPK
	SEIAHRYNDLGEQHFK
	TNCDLYEKLGEYGFQNAILVR
	VCLLHEKTPVSEHVTK
	YNDLGEQHFKGLVLIAFSQYLQK

Note: The identified peptides shared in both SDS-PAGE-resolved avidin-enriched sample (Sample 1) and 2-DE-resolved unenriched sample (Sample 2) are labeled in red.

Table 2.3 Details of 29 shared proteins identified in both samples by MaxQuant.

UniProt Entry	Protein Names	Gene Names	Molecular Weight (Da)
P01027	Complement C3 (HSE-MSF) [Cleaved into: Complement C3 beta chain; C3-beta-c (C3bc); Complement C3 alpha chain; C3a anaphylatoxin; Acylation stimulating protein (ASP) (C3adesArg); Complement C3b alpha' chain; Complement C3c alpha' chain fragment 1; Complement C3dg fragment; Complement C3g fragment; Complement C3d fragment; Complement C3f fragment; Complement C3c alpha' chain fragment 2]	C3	186,484
P01942	Hemoglobin subunit alpha (Alpha-globin) (Hemoglobin alpha chain)	Hba Hba-a1	15,085
P02088	Hemoglobin subunit beta-1 (Beta-1-globin) (Hemoglobin beta-1 chain) (Hemoglobin beta-major chain)	Hbb-b1	15,840
P05213	Tubulin alpha-1B chain (Alpha-tubulin 2) (Alpha-tubulin isotype M-alpha-2) (Tubulin alpha-2 chain) [Cleaved into: Detyrosinated tubulin alpha-1B chain]	Tuba1b Tuba2	50,152
P07724	Serum albumin	Alb Alb-1 Alb1	68,693
P07759	Serine protease inhibitor A3K (Serpin A3K) (Contrapsin) (SPI-2)	Serpina3k Mcm2 Spi2	46,880
P07901	Heat shock protein HSP 90-alpha (Heat shock 86 kDa) (HSP 86) (HSP86) (Tumor-specific transplantation 86 kDa antigen) (TSTA)	Hsp90aa1 Hsp86 Hsp86-1 Hspca	84,788
P11499	Heat shock protein HSP 90-beta (Heat shock 84 kDa) (HSP 84) (HSP84) (Tumor-specific transplantation 84 kDa antigen) (TSTA)	Hsp90ab1 Hsp84 Hsp84-1 Hspcb	83,281
P14824	Annexin A6 (67 kDa calelectrin) (Annexin VI) (Annexin-6) (Calphobindin-II) (CPB-II) (Chromobindin-20) (Lipocortin VI) (Protein III) (p68) (p70)	Anxa6 Anx6	75,885
P20029	Endoplasmic reticulum chaperone BiP (EC 3.6.4.10) (78 kDa glucose-regulated protein) (GRP-78) (Binding-immunoglobulin protein) (BiP) (Heat shock protein 70 family protein 5) (HSP70 family protein 5) (Heat shock protein family A member 5) (Immunoglobulin heavy chain-binding protein)	Hspa5 Grp78	72,422
P23953	Carboxylesterase 1C (EC 3.1.1.1) (Liver carboxylesterase N) (Lung surfactant convertase) (PES-N)	Ces1c Es1	61,056
P40142	Transketolase (TK) (EC 2.2.1.1) (P68)	Tkt	67,630
P50446	Keratin, type II cytoskeletal 6A (Cytokeratin-6A) (CK-6A) (Keratin-6-alpha) (mK6-alpha) (Keratin-6A) (K6A)	Krt6a Ker2 Krt2-6 Krt2-6a Krt6	59,335
P52480	Pyruvate kinase PKM (EC 2.7.1.40) (Pyruvate kinase muscle isozyme)	Pkm Pk3 Pkm2 Pykm	57,845

P60710	Actin, cytoplasmic 1 (Beta-actin) [Cleaved into: Actin, cytoplasmic 1, N-terminally processed]	Actb	41,737
P62737	Actin, aortic smooth muscle (Alpha-actin-2) [Cleaved into: Actin, aortic smooth muscle, intermediate form]	Acta2 Actsa Actvs	42,009
P63017	Heat shock cognate 71 kDa protein (Heat shock 70 kDa protein 8)	Hspa8 Hsc70 Hsc73	70,871
P63260	Actin, cytoplasmic 2 (Gamma-actin) [Cleaved into: Actin, cytoplasmic 2, N-terminally processed]	Actg1 Actg	41,793
P63268	Actin, gamma-enteric smooth muscle (Alpha-actin-3) (Gamma-2-actin) (Smooth muscle gamma-actin) [Cleaved into: Actin, gamma-enteric smooth muscle, intermediate form]	Actg2 Acta3 Actsg	41,877
P68033	Actin, alpha cardiac muscle 1 (Alpha-cardiac actin) [Cleaved into: Actin, alpha cardiac muscle 1, intermediate form]	Actc1 Actc	42,019
P68369	Tubulin alpha-1A chain (Alpha-tubulin 1) (Alpha-tubulin isotype M-alpha-1) (Tubulin alpha-1 chain) [Cleaved into: Detyrosinated tubulin alpha-1A chain]	Tuba1a Tuba1	50,136
P68373	Tubulin alpha-1C chain (Alpha-tubulin 6) (Alpha-tubulin isotype M-alpha-6) (Tubulin alpha-6 chain) [Cleaved into: Detyrosinated tubulin alpha-1C chain]	Tuba1c Tuba6	49,909
Q6NXH9	Keratin, type II cytoskeletal 73 (Cytokeratin-73) (CK-73) (Keratin-73) (K73) (Type II inner root sheath-specific keratin-K6irs3) (Type-II keratin Kb36)	Krt73 Kb36	58,911
Q8BFZ3	Beta-actin-like protein 2 (Kappa-actin)	Actbl2	42,004
Q8BGZ7	Keratin, type II cytoskeletal 75 (Cytokeratin-75) (CK-75) (Keratin-6 hair follicle) (mK6hf) (Keratin-75) (K75) (Type II keratin-K6hf) (Type-II keratin Kb18)	Krt75 Kb18	59,741
Q8VED5	Keratin, type II cytoskeletal 79 (Cytokeratin-79) (CK-79) (Keratin-79) (K79) (Type-II keratin Kb38)	Krt79 Kb38	57,552
Q91X72	Hemopexin	Hpx Hpxn	51,318
Q922U2	Keratin, type II cytoskeletal 5 (Cytokeratin-5) (CK-5) (Keratin-5) (K5) (Type-II keratin Kb5)	Krt5 Krt2-5	61,767
Q9Z331	Keratin, type II cytoskeletal 6B (Cytokeratin-6B) (CK-6B) (Keratin-6-beta) (mK6-beta) (Keratin-6B) (K6B)	Krt6b K6-beta Krt2-6b	60,322

Note: 29 proteins were identified in both samples (SDS-PAGE-resolved avidin-enriched sample and 2-DE-resolved unenriched sample) by MaxQuant.

Table 2.4 Peptides and their corresponding proteins identified by MODa.

Identified Peptides from Sample 1	Corresponding Sample 1 Proteins	Identified Peptides from Sample 2	Corresponding Sample 2 Proteins
DVFLGTFLYEYSR	P07724	DVFLGTFLYEYSR	P07724
ENPTTFMGHYLHEVAR	P07724	ENPTTFMGHYLHEVAR	P07724
GLVLIAFSQYLQK	P07724	GLVLIAFSQYLQK	P07724
LGEYGFQNAILVR	P07724	LGEYGFQNAILVR	P07724
QTALAELVK	P07724	QTALAELVK	P07724
FLEQQNQVLQTK	Q6IFZ6	FLEQQNQVLQTK	Q6IFZ6
VNSDEVGGEALGR	P02088	APQVSTPTLVEAAR	P07724
DLVVLLFETALLSSGFSLEDPQTHSNR	P11499	CCSGSLVER	P07724
GTGASGSFK	Q07133	DTCFSTEGPNLVTR	P07724
GFSSGSAVVSGGSR	Q3TTY5	HPDYSVSLLLR	P07724
SADELENLILQQN	Q6DIB4	LPCVEDYLSAILNR	P07724
SGSEGPVLLLLHGGGHSALSWAVFTAAIISR	Q8BVQ5	LSQTFPNADFAEITK	P07724
SSAFDGLLPQQN	Q8VBT3	LVQEVTDFAK	P07724
GVTHNIPLLR	Q91ZA3	PCFSALTVDETYVPK	P07724
HIEIQVLGDK	Q91ZA3	SLHTLFGDK	P07724
LVTYGSDR	Q91ZA3	TPVSEHVTK	P07724
SFGLPSIGR	Q91ZA3	YNDLGEQHFK	P07724
VVEEAPSIFLDPETR	Q91ZA3	DIPVDSPELK	O08677
YSSAGTVEFLVDSQK	Q91ZA3	ALLNVVDSAR	O35405
LYLGHNYVTAIR	Q921I1	IGGHGAEYGAEALER	P01942
SDIVIAVTYNR	Q99MR8	VITAFNDGLNHLDSLK	P02088
VFFSEGAQANR	Q99MR8	YFDSFGDLSSASAIMGNAK	P02088
LGLALNFSVFYYEILNSPEK	Q9CQV8	LHVDPENFR	P02089
SLQSVAEER	Q9CZM2	VVAGVATALAHK	P02089
NAFASVILFGTNNSSSISGVWVFR	Q9D8N0	LAQIHFPR	P07758
		DLQILAEFHEK	P07759
		ELISELDER	P07759
		EVFTEQADLSGITETK	P07759
		LSVSQVVHK	P07759
		TMEEILEGLK	P07759
		EADDIVNWLK	P09103
		THILLFLPK	P09103
		VDATEESDLAQQYGVR	P09103
		ALELTGLK	P09405
		ETLEEVFEK	P09405
		FAISELFAK	P09405
		GFGFVDFNSEEDAK	P09405

		TLVLSNLSYSATK	P09405
		ATGVFTTLQPLR	P11276
		GLTPGVIYEGQLISIQQYGHR	P11276
		HFSVEGQLEFR	P11499
		HLEINPDHPIVETLR	P11499
		SLTNDWEDHLAVK	P11499
		DAFVAIVQSVK	P14824
		DLESDIIGDTSGHFQK	P14824
		ENDDVVSEDLVQQDVQDLYEAGELK	P14824
		GIGTDEATIIDIVTHR	P14824
		GSVHDFPEFDANQDAEALYTAMK	P14824
		ILISLATGNR	P14824
		LVFDEYLK	P14824
		SEIDLLNIR	P14824
		SELDMLDIR	P14824
		TLIEILATR	P14824
		IINEPTAAAIAYGLDK	P16627
		FEELNADLFR	P17156
		LLQDFFNGK	P17156
		DNHLLGTFDLTGIPPAPR	P20029
		ELEEIVQPIISK	P20029
		IEWLESHQDADIEDFK	P20029
		NELESYAYSLK	P20029
		SDIDEIVLVGGSTR	P20029
		SQIFSTASDNQPTVTIK	P20029
		IFNNGADLSGITEENAPLK	P22599
		LVQIHIPR	P22599
		SFNTVPYIVGFNK	P23953
		NPDTNIVFSPLSISAALAIVSLGAK	P29621
		FATNFYQHLADSK	P32261
		TSDQIHFFFAK	P32261
		FAEAFEAIPR	P42932
		LATNAAVTVLR	P42932
		LFVTNDAATILR	P42932
		VADIALHYANK	P42932
		TLEAQLTPQVVER	P49182
		YEVTTIHNLFR	P49182
		FTGLQYLR	P51885
		SLEYLDLSFNQMSK	P51885
		GVNLPGAAVDLPAVSEK	P52480
		IYVDDGLISLQVK	P52480

		VGWEQLLTTIAR	P57780
		VGLQVVAVK	P63038
		SPYQLVLQHSR	P82198
		ISHLPLVEELR	P97310
		GILLYGPPGTGK	Q01853
		LDQLIYIPLPDEK	Q01853
		LSQELDFVSHNVR	Q02819
		DTPLTLTVLHK	Q06770
		GFLDVVTR	Q3TW96
		VQDPLAELVK	Q497V5
		SYIHEVAR	Q5SYD0
		GSVHQNFDDFTFVTGK	Q60963
		TLQPLLFINSAK	Q60963
		AYYHLLEQVAPK	Q61233
		EGESLEDLMK	Q61233
		LNLAFIANLFNK	Q61233
		NEALIALLR	Q61233
		NWMNSLGVNPR	Q61233
		QFVTATDVVR	Q61233
		TLTLALVWQLMR	Q61233
		VYALPEDLVEVNPK	Q61233
		YAFVNWINK	Q61233
		DFFHLDER	Q61247
		QEEDLANINQWVK	Q61247
		GPLAHQISGLFLPSK	Q61503
		AISHEHSPSDLEAHFVPLVK	Q76MZ3
		LAGGDWFTSR	Q76MZ3
		SEIIPMFSNLASDEQDSVR	Q76MZ3
		VLELDNVK	Q76MZ3
		EIISEVQR	Q7SIG6
		LHFFMPGFAPLTSR	Q7TMM9
		SHIDQLVLIFAGK	Q8R317
		QTPTFWILAR	Q8VBW6
		CSPDPGLTALLSDHR	Q91X72
		FNPVTGEVPPR	Q91X72
		GATYAFTGSHYWR	Q91X72
		GPDSVFLIK	Q91X72
		WFWDFATR	Q91X72
		ANDDIIVNWVNR	Q99K51
		AYFHLLNQIAPK	Q99K51
		HVSPAGAAVGVPLSEDEAR	Q9CWJ9
		SLASLGLSLVASGGTAK	Q9CWJ9
		LQLLNLSR	Q9DBB9

		APLVPPGSPVVNALFR	Q9JHU9
		SVLVDFLIGSGLK	Q9JHU9
		QHFEWLLK	Q9QUR6
		LVVLPFPGK	Q9QXC1

Note: AlbM peptides are highlighted in yellow. Shared peptides and their corresponding proteins (shown as UniProt accession numbers), are labeled in red. Sample 1 is the SDS-PAGE-resolved avidin-enriched sample. Sample 2 is the 2-DE-resolved unenriched sample.

Table 2.5 Venn data table showing relations among the identification results.

(0001)	*(0010)*	*(0100)*	*(0101)*	*(1000)*	*(1010)*	*(1011)*	*(1100)*	*(1110)*	*(1111)*
Q07133	O35405	A0JP43	Q3TTY5	A2ASS6	O08677	Q6IFZ6	P01027	P01942	P02088
Q6DIB4	P16627	E9Q286	Q8BVQ5	A6X935	P02089		P05213	P07759	P07724
Q8VBT3	P29621	P01872	Q91ZA3	C8YR32	P07758		P07901	P14824	P11499
Q9CZM2	P97310	P02301	Q921I1	O35969	P09103		P40142	P20029	
	Q497V5	P02535	Q99MR8	O55055	P09405		P50446	P23953	
	Q5SYD0	P06151	Q9CQV8	O70318	P11276		P60710	P52480	
	Q7SIG6	P07744	Q9D8N0	O88491	P17156		P62737	Q91X72	
		P08071		O89110	P22599		P63017		
		P0C6F1		P04104	P32261		P63260		
		P10126		P04441	P42932		P63268		
		P12382		P05214	P49182		P68033		
		P14148		P08003	P51885		P68369		
		P15864		P0CG49	P57780		P68373		
		P26041		P0CG50	P63038		Q6NXH9		
		P35700		P10077	P82198		Q8BFZ3		
		P35980		P10711	Q01853		Q8BGZ7		
		P43274		P16045	Q02819		Q8VED5		
		P43276		P16301	Q06770		Q922U2		
		P43277		P17182	Q3TW96		Q9Z331		
		P46662		P18529	Q60963				
		P47857		P20152	Q61233				
		P47911		P20918	Q61247				
		P47915		P21614	Q61503				
		P47962		P24527	Q76MZ3				
		P49962		P27005	Q7TMM9				
		P53026		P29699	Q8R317				
		P54310		P42859	Q8VBW6				
		P62631		P48036	Q99K51				
		P62806		P59240	Q9CWJ9				

		P62849		P61979	Q9DBB9				
		P62855		P62983	Q9JHU9				
		P62918		P62984	Q9QUR6				
		P63101		P68134	Q9QXC1				
		P68433		P68368					
		P84228		P68372					
		P84244		P80316					
		P86048		P83626					
		Q05920		P97386					
		Q3USH1		P97504					
		Q3UV17		P99024					
		Q5DTX6		Q00896					
		Q5SWU9		Q00897					
		Q61881		Q00PI9					
		Q6ZWV3		Q01514					
		Q8BMK4		Q03526					
		Q8BQM9		Q3TLH4					
		Q8BRH4		Q3UMC0					
		Q8BTI8		Q3UTQ7					
		Q8C341		Q3UXZ6					
		Q8VEK3		Q4VA45					
		Q91YQ5		Q569L8					
		Q9CR57		Q60674					
		Q9CTN5		Q61316					
		Q9CXB8		Q61329					
		Q9CY62		Q61576					
		Q9D8E6		Q61702					
		Q9DBN5		Q62255					
		Q9DCV7		Q64438					
		Q9QYB2		Q684R7					
				Q6IFZ9					

				Q6P8J2					
				Q6P9J9					
				Q7TNG5					
				Q7TNP2					
				Q7TPR4					
				Q80TK0					
				Q80X19					
				Q80YV3					
				Q811F1					
				Q8BGA8					
				Q8BGC4					
				Q8BKI2					
				Q8BND5					
				Q8BTF7					
				Q8BTY8					
				Q8BVK9					
				Q8C145					
				Q8C166					
				Q8C551					
				Q8C5W4					
				Q8K0D2					
				Q8K3V4					
				Q8R0W0					
				Q8R2S9					
				Q8VCT9					
				Q8VDC1					
				Q8VDW0					
				Q8VIM9					
				Q91WB4					
				Q91WG0					
				Q99J45					
				Q99K95					
				Q99KC8					
				Q99KD5					
				Q99KW3					
				Q9CQB5					

				Q9CQW3					
				Q9CWF2					
				Q9CXV9					
				Q9D0F9					
				Q9D281					
				Q9D3R6					
				Q9D4B1					
				Q9D554					
				Q9D6F9					
				Q9DC60					
				Q9EP96					
				Q9ERC8					
				Q9ERD7					
				Q9JJN2					
				Q9JKF1					
				Q9JLC8					
				Q9JMA1					
				Q9JMH6					
				Q9R0H5					
				Q9WU79					
				Q9WUP0					
				Q9WVE8					
				Q9Z0G7					
				Q9Z1N5					
				Q9Z1X4					
				Q9Z247					

Note: This Venn data table shows detailed identification results of both SDS-PAGE-resolved avidin-enriched sample (Sample 1) and 2-DE-resolved unenriched sample (Sample 2) by both MODa and MaxQuant. Column titles refer to the specific areas on the Venn diagram above in **Figure 2.5**.

Table 2.6 Homology analysis of serum albumin from multiple species showing favored conservation of the inner peptide.

Species	Albumin ID	N-terminus …	Sequence region 1 (showing the inner peptide)	Omitted region (…)	Sequence region 2 (showing the surface peptide)	C-terminus
SALSA	Q03156	… 459	CCKDEPGHFVLPCAEEKLTDAIDATCDDYDPSSINPHIAHCCNQSYSMRR	508 … 580	KCCAAEDQAA FTEEAPKLVSESAELVKV--	608
SALSA	P21848	… 459	CCKDEQGHFVLPCAEEKLTDAIDATCDDYDPSSINPHIAHCCNQSYSMRR	508 … 580	KCCAAEDQAA FTEEAPKLVSESAELVKV--	608
CHICK	P19121	… 465	CCQ-LGEDRRMACSEGYLSIVIHDTCRKQETTPINDNVSQCCSQLYANRR	513 … 585	KCCKQSDINT FGEEGANLIVQSRATLGIGA	615
MERUN	O35090	… 462	CCA-LPEKKRLPCVEDYLSAILNRVCLLHEKTPVSEQVTKCCSGSLVERR	510 … 580	KCCKQEDKEA FSTEGPKLVAESQKALA---	609
MESAU	A6YF56	… 461	CCV-LPEAQRLPCVEDYISAILNRVCVLHEKTPVSEQVTKCCTGSVVERR	509 … 581	KCCKAEDKEA FSEDGPKLVASSQAALA---	608
MOUSE	P07724	… 461	CCT-LPEDQRLPCVEDYLSAILNRVCLLHEKTPVSEHVTKCCSGSLVERR	509 … 581	TCCKAADKDT FSTEGPNLVTRCKDALA---	608
RAT	P02770	… 461	CCT-LPEAQRLPCVEDYLSAILNRLCVLHEKTPVSEKVTKCCSGSLVERR	509 … 581	KCCKAADKDN FATEGPNLVARSKEALA---	608
RABIT	P49065	… 461	CCK-HPEAERLPCVEDYLSVVLNRLCVLHEKTPVSEKVTKCCSESLVDRR	509 … 581	KCCSAEDKEA FAVEGPKLVESSKATLG---	608
PIG	P08835	… 460	CCK-RPEEERLSCAEDYLSLVLNRLCVLHEKTPVSEKVTKCCTESLVNRR	508 … 580	KCCAAPDHEA FAVEGPKFVIEIRGILA---	607
BOVIN	P02769	… 460	CCT-KPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTESLVNRR	508 … 580	KCCAADDKEA FAVEGPKLVVSTQTALA---	607
SHEEP	P14639	… 460	CCA-KPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTESLVNRR	508 … 580	KCCAADDKEG FVLEGPKLVASTQAALA---	607
EQUAS	Q5XLE4	… 460	CCK-LPESERLPCSENHLALALNRLCVLHEKTPVSEKITKCCTDSLAERR	508 … 580	KCCGAEDKEA FAEEGPKLVASSQLALA---	607
HORSE	P35747	… 460	CCK-LPESERLPCSENHLALALNRLCVLHEKTPVSEKITKCCTDSLAERR	508 … 580	KCCGREDKEA FAEEGPKLVASSQLALA---	607
MACFA	A2V9Z4	… 461	CCK-LPEAKRMPCAEDYLSVVLNRLCVLHEKTPVSEKVTKCCTESLVNRR	509 … 581	KCCKADDKEA FAEEGPKFVAASQAALA---	608
MACMU	Q28522	… 453	CCK-LPEAKRMPCAEDYLSVVLNRLCVLHEKTPVSEKVTKCCTESLVNRR	501 … 573	KCCKADDKEA FAEEGPKFVAASQAALA---	600
HUMAN	P02768	… 461	CCK-HPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRR	509 … 581	KCCKADDKET FAEEGKKLVAASQAALGL--	609
PONAB	Q5NVH5	… 461	CCK-HPEPKRMPCAEDYLSVVLNQLCVLHEKTPVSERVTKCCTESLVNRR	509 … 581	KCCKADDKET FAEEGKKLVAASQAALGL--	609
FELCA	P49064	… 461	CCT-HPEAERLSCAEDYLSVVLNRLCVLHEKTPVSERVTKCCTESLVNRR	509 … 581	KCCAAEDKEA FAEEGPKLVAAAQAALA---	608
CANLF	P49822	… 461	CCK-KPESERMSCAEDFLSVVLNRLCVLHEKTPVSERVTKCCSESLVNRR	509 … 581	KCCAAENKEG FSEEGPKLVAAAQAALV---	608
BOMMX	Q3T478	… 460	CCA-LPNTQKMPCAEGGLSLIIGEFCEMEKTHPINEHVKNCCWKSYSNRR	508 … 580	KCCAAEDHQA FNAEEPILIEHCKQLAA---	607
XENLA	P14872	… 460	CCA-VPENQRMPCAEGDLTILIGKMCERQKKTFINNHVAHCCTDSYSGMR	508 … 580	KCCTADEHQP FDTEKPVLIEHCQKLHP---	607
XENLA	P08759	… 459	CCA-VPENQRMPCAEGDLTILIGKMCERQKKTFINNHVAHCCTDSYSGMR	507 … 579	KCCTADEHQP FDTEKPVLIEHCQKLHP---	606
	Output		** : * * :: : * . :. .: :** *		.** : ** : ::	

Legend.

Alignment output symbol	Indication
*	Fully conserved residue positions
:	Strongly conserved residue positions (shared by groups of amino acids with similar chemical properties)
.	Weakly conserved residue positions (shared by groups of amino acids with slightly distinct chemical properties)

Note: The *in silico* tryptic peptides are highlighted. Note the intact-probe-modified peptide in the deep pocket (yellow) contains more conserved residues than reduced-probe-modified peptide on the AlbM surface (turquoise). The cysteine residues (in red and orange) of both peptides are fully conserved. The sequence alignment was performed with a popular algorithm Clustal Omega.[196]

Chapter 3 Measuring Proteome-wide Live-cell Actions of Small Molecules Using α-Methylene-β-lactone and Mass Spectrometry

This chapter reports novel utilities of the α-methylene-β-lactone (MeLac) moiety as a chemical probe warhead of multiple electrophilic sites. This study demonstrates that MeLac-alkyne is a competent covalent probe and reacts with diverse proteins in live cells. Proteomics analysis of affinity-enriched samples identified probe-reacted proteins, resolved their modified peptides/residues, and thus characterized probe-protein reactions. Unique methods have been developed to evaluate confidence in identification of the reacted proteins and modified peptides. Tandem mass spectra of the peptides have uncovered that MeLac reacts with nucleophilic cysteine, serine, lysine, threonine, and tyrosine residues, through either Michael addition or acyl addition. As a broad-spectrum measurement probe, MeLac-alkyne has successfully analyzed orlistat and parthenolide selectivity in live HT-29 cells within a unique peptide-centric proteomics platform. As a scaffold, MeLac-alkyne has created a selective β-lactone probe with glutathione via proteome-assisted probe-ligand assembly in biological matrices. The MeLac-alkyne-empowered chemical proteomics platform is widely adaptable for measuring the live-cell action of reactive molecules. The assembly of MeLac-alkyne glutathione adduct exemplifies a scalable route to develop selective probes. It also provides an unprecedented opportunity to probe the glutathione S-transferase P1 (GSTP1) responsible for multi-drug resistance in cancer patients. Overall, MeLac is a versatile warhead bearing enormous potential in expediting the development of chemical probes and targeted covalent inhibitors. MeLac-based probes are novel chemical tools making easier interrogating protein (re)activity and developing new medicine.

Part of this study has been published on ChemRxiv under the title "α-Methylene-β-Lactone Probe for Measuring Live-Cell Reactions of Small Molecules" as a preprint.[197]

3.1 Introduction

3.1.1 Significance of chemical proteomics in drug development

Bioactive small molecules have always been the focus of interest for developing diagnostic and therapeutic agents to elucidate and alter biochemical pathways. Drug discovery pipelines typically involve two approaches for screening bioactive small molecules: target-based screening, where compounds are developed to target a disease-causing enzyme, and phenotypic screening, where libraries of small molecules are screened against model organisms to revert a disease-related phenotype.[198,199] In general, the target-based approach, which operates at a lower level, is target-specific, thus reliable and low-risk. However, the implementation of target-based screening is usually restricted by the progress of biological understanding of the disease and throttled by the unforeseeable ADME-Tox (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profile. On the other hand, phenotypic screening, which operates at a higher level and factors in drug delivery and toxicity, rarely provides opportunities to discover and validate the mode of action of a drug candidate.[200] Although expeditable and scalable, phenotypic screening is always riskier to perform due to the fact that chemically distinct drug candidates may function on different modes of action in different contexts or organisms beyond the experimental expectation. Fortunately, the fast-advancing field of chemical proteomics provides a suitable middle-ground between target-based and phenotypic screening approaches for unbiased exploration of drug-target-phenotype relationships.[201]

Chemical proteomics is a progressive subfield of chemical biology. Chemical proteomics tactically integrates synthetic chemistry to generate investigational small molecules to study and manipulate whole sets of proteins, known as proteomes, and bioanalytical chemistry to establish

bioanalytical platforms. These platforms investigate modes of action for bioactive small molecules and functions of proteins.[74] Chemical proteomics analysis of drug-protein reactions expedites the development of new drugs by revealing selective inhibitors early, identifying toxicity liabilities, and mitigating the risk of late-stage failures.[74] In a typical chemical proteomics workflow, direct chemical perturbation is usually introduced as a measure of treatment to a model proteome at a specific level of biological relevance, such as cell lysates, living cells, tissues, and animal specimens. Afterward, latent measurements of biochemical changes are performed. Offering multiplexed analyses on, these chemical proteomics platforms often facilitate quantitative studies of target engagement, off-target effect, and cytotoxicity for a compound of interest simultaneously.[202-204]

3.1.2 Chemical probes and chemical proteomics methodologies

The measurement of post-treatment biochemical changes of a proteome relies on an elaborate set of analytical and chemical tools. The analytical tools can be either optics-based or mass spectrometry-based. The chemical probes are either covalent or non-covalent. Although affinity capture approaches for target protein profiling are used with both types of chemical probes, each type of chemical probes denotes a distinguishable analytical and biochemical methodology of the underlying chemical proteomics experiments. Compound-centric approaches feature chemically modified, either tagged[71] or immobilized[72], compounds of interest as either covalent or non-covalent chemical probes. Also known as boutique probes, these chemical probes are introduced directly as baits to capture target proteins from a complex proteome. In contrast, activity-based approaches depend on meticulously designed covalent chemical probes. These activity-based probes are capable of irreversibly binding, reacting with, and labeling target proteins from a complex proteome, which is pre-treated with compounds of interests. The subsequent

elucidation of target activity relies on the indirect measurement of probe-target adducts as competitive binding assays.

Presumably, compound-centric chemical proteomics is the most straightforward strategy for studying the proteome-wide activity of a known drug molecule. It mandates a boutique probe created by installing a reporting group on the drug molecule. However, the construction of a drug-derived probe often bears high synthetic costs. The application of such a boutique probe also suffers from the consequence of chemical modifications. Because any structural changes made to the drug molecule may significantly alter the parent molecule's potency and selectivity profile.[73] The more sophisticated strategy is activity-based chemical proteomics, which emphasizes the use of a broad-spectrum covalent probe. Activity-based chemical proteomics, technically referring to competitive ABPP, offers a nearly universal bioanalytical platform.[74,75] Such a platform can effectively decipher the proteome-wide action of underivatized drugs, as well as other reactive molecules like environmental toxins and reactive metabolites from the human microbiota.[76-79]

In contrast to boutique probes, ABPs are designed to carry less specificity to their protein targets or are not specific at all. They capture proteome-wide "snapshots" visualizing drug-protein interactions by permanently occupying available active sites post drug treatment on the model proteome. As illustrated in **Scheme 3.1**, a competitive ABPP platform depends on its ABP to measure the proteome-wide action of an underivatized drug.[80,81] The distinctive technological advantage of competitive ABPP is that a single ABP with broad proteome coverage can establish a versatile analytical platform capable of evaluating multiple drugs or drug candidates on different subsets of a single proteome. Therefore, ABPs with a broad spectrum of reactivity can fully unleash the enormous potential of competitive ABPP technology.

An ABP normally consists of three chemical components: a reactive group that is commonly known as the "warhead," a recognizing group that is commonly established as a chemically inert binding moiety, and a reporting group that is commonly devised as an affinity/fluorophore tag. (top, **Scheme 3.1**) An ABP's warhead dictates the probe's reactivity while the recognizing group mainly determines the probe's specificity towards a subset of proteins within a proteome. In principle, when a broad reactivity spectrum is desired, an ABP should incorporate (1) no recognizing group that may direct pre-reaction probe-protein complexation and (2) a warhead of broad reactivity spectrum. While fluorophosphonate-based probes provide a successful example of exploiting the first option of this principle,[205] the second option is rarely explored due to the limited availability of broad-reactivity warhead. It has always been a challenge to design a warhead that covers the diverse reactivity space of a proteome. Individual proteins differ in molecular composition and structure. Their reactions with small molecules are distinct from site to site, domain to domain, and protein to protein. The diversity of these reactions rapidly multiplies due to the large number and abundance range of proteins in a proteome.

3.1.3 The emerging need for novel warheads of broad reactivity

Existing warheads are far from ideal for universal platforms of competitive ABPP. To date, a diverse collection of reactivity of warheads has been implemented on ABPs, including Michael addition, non-Michael nucleophilic addition, addition-elimination, nucleophilic substitution, and oxidation.[124,206,207] Unfortunately, these conventional warheads typically react with only one functional group on amino acid residues: cysteine,[119,208,209] serine,[205,210,211] threonine, tyrosine,[212] lysine,[79,212,213] and histidine,[212] resulting in limited proteome coverage. Their reactivity depends on either a specific reaction mechanism or a single electrophilic site.[79,122-124,214-216] After all, a widely adaptable platform of competitive ABPP requires a versatile probe equipped with a

warhead of broad reactivity. To broaden the reactivity spectrum further, ABPs should adopt a warhead with multiple electrophilic sites, which can extend the coverage of diverse functional groups on amino acid residues. Nonetheless, due to the lack of warheads consisting of multiple electrophiles, such an ideal broad-spectrum ABP is still absent.

This study employed α-methylene-β-lactone (MeLac) as a novel warhead with broad reactivity and live-cell compatibility. Small and rigid, MeLac couples the Michael acceptor functionality of acrylate[217,218] with β-lactone reactivity.[219-221] MeLac was proposed to be susceptible to nucleophilic attacks at three distinct sites, as shown on **Scheme 3.2**, which would result in a broad reactivity spectrum covering multiple nucleophilic side chains on different amino acid residues. Therefore, this hypothesis was tested as follows: when used on a chemical probe as the warhead for proteomics profiling, MeLac would react with different nucleophilic groups on proteins via different mechanisms. This new warhead is prone to reactions with nucleophilic thiol (Cys), hydroxyl (Ser, Thr, and Tyr), and amino (Lys) groups on the protein. Also, MeLac can provide separate sites for regioselective reactions with spatially arranged protein nucleophiles. Thus, the reactivity of the small and rigid MeLac is projected to be broad. A MeLac-equipped chemical probe would make a competent measurement probe for building a widely adaptable platform of competitive ABPP.

Using a MeLac-alkyne as a probe without recognizing group, the probe-protein reactions in live cells were examined and characterized using both gel-based and mass spectrometry-based proteomics profiling. New methods for the confident identification of protein adducts were developed. A peptide-centric live cell compatible platform of competitive ABPP was built using MeLac-alkyne. This platform was used for analyzing the selectivity profiles of protein reactions with three chemically distinct compounds (**Scheme 3.3**): orlistat, parthenolide, and alkyl MeLac,

which were used as model inhibitors with the β-lactone, Michael acceptor, and MeLac reactivity, respectively.

3.1.4 The versatility of a chemical scaffold

On the other hand, MeLac was also introduced as a useful scaffold in developing selective probes. MeLac offers separate electrophilic sites allowing not only regioselective reactions with spatially arranged protein nucleophiles but also scalable development of selective probes to protein targets of interest. Derivatization of one electrophilic site on the MeLac scaffold with an affinity ligand can introduce a recognizing group *in-situ* for its pre-reaction complexation with the active site on a protein. In this study, the in-cell conjugation of reduced glutathione (GSH) to MeLac was also discovered. This surprising intracellular process produced a highly selective β-lactone boutique probe for GSH. Overall, this study demonstrates utilities of the MeLac scaffold in developing (1) broad-reactivity ABPs for measuring proteome-wide actions of underivatized small molecules and (2) selective boutique probes for specific affinity ligands targeting small groups of proteins.

3.2 Experimental

3.2.1 Overview

To investigate the intriguing broad reactivity aspect of MeLac, the 4-(but-3-ynyl)-3-methyleneoxetan-2-one (MeLac-alkyne probe, **Scheme 3.3**)[222,223] was studied for its protein target and labeling site-specific profile. After reactivity and target characterization of MeLac-alkyne, it was then implemented as a competitive ABP for analyzing proteome-wide reactions of three compounds in live HT-29 cells. These three compounds were 4-decyl-4-methyleneoxetan-2-one (alkyl MeLac inhibitor), orlistat, and parthenolide. (**Scheme 3.3**) Both MeLac-alkyne and alkyl MeLac inhibitor were synthesized and characterized at Howell group at Department of Chemistry. Orlistat and parthenolide were purchased commercially. MeLac-alkyne probe-protein reactions were initially evaluated using gel-based ABPP. Next, as illustrated in **Scheme 3.4**, probe-reacted proteins were conjugated to azide-(desthio)biotin tags and enriched with immobilized avidin. This affinity enrichment allowed the selective release of probe-reacted proteins and probe-modified peptides for the following MS-based probe reactivity characterization, target profiling, structural elucidation of probe-modified peptides, and dose-dependent target site-specific profiling of benchmark inhibitors using MeLac-alkyne in competitive ABPP.

3.2.2 HT-29 cell culture

HT-29 cells were cultured to a confluency of 90-95%, as shown on **Figure 3.1** in either T75 or T175 cell culture flasks (CELLTREAT Scientific; about 8 million cells/ T75 flask or 23 million cells/T175 flask) with appropriate amount of growth medium (GM). The GM was prepared by supplementing Dulbecco's Modified Eagle Medium (DMEM, Gibco) with 10% (v/v) fetal

bovine serum (FBS, Gibco) and 1% (v/v) Pen/Strep antibiotics (FBS, Gibco). Sterile disposable filter units (with a pore size of 0.22 μm or 0.45 μm) were used to sterilize the GM before its use for cell culture. The incubation of cell culture flasks was performed in a humidified incubator at 5% of CO_2 and 37 °C.

To initiate the culturing of HT-29 cells, a vial of cryogenically preserved cells (seeds) was thawed and washed with GM at 37 °C. The washed cells were then transferred to a proper cell culture flask at a proper seeding density, as suggested in **Table 3.1**. GM change (replacing old GM with fresh GM) was performed three to four times per week, depending on the cell growth and viability. To maintain a healthy HT-29 cell line, cells were routinely passed at a confluency of above 75%. For each passage, cells were detached from the supporting surface of the vessel by trypsin/EDTA solution (Gibco). The detached culture suspension was then split at a ratio of 1:3 or 1:4 into new vessels of the same size. Cell viability (the number of living cells divided by the number of total cells) was checked on a small aliquot of the culture suspension treated with trypan blue on a hemocytometer under a microscope.

3.2.3 *In vitro* probe-proteome reaction for competitive ABPP using MeLac-alkyne probe and HT-29 cells

Before reaction, cells were washed twice with 10 mL (for cells in each T75 flask) or 20 mL (for cells in each T175 flask) of phosphate-buffered saline (PBS, 1X, pre-heated at 37 °C, Gibco).

For inhibitor-free probe treatment of the cells, either 10 or 20 μL of a pre-aliquoted stock solution of MeLac-alkyne probe (dissolved in dimethylacetamide, DMAC, Sigma-Aldrich) or its diluate in situ was added to each flask containing pre-washed cells and 10 or 20 mL of PBS, resulting in working probe concentration of 0.01, 0.05, 0.1, 0.5, 1, 5, 10, or 50 μM. To incubate

cells, the flasks were placed on a rocking platform at 37 °C for 30 minutes. After the reaction, 10 or 20 mL of ice-cold tris-buffered saline (TBS, 1X, pre-heated at 37 °C, Pierce) was added to each flask to quench the reaction. Cells in each flask were then gently scraped off the flask surface. The subsequent suspension was transferred to either a 15-mL or 50-mL conical tube, centrifuged at 500 xg, 4 °C for 3 min to pellet the cells. Cells in each conical tube were washed with 10 or 20 mL of PBS. Cell pellets were either used immediately or stored at -80 °C.

For the competitive activity-based proteomics profiling, the inhibitor (orlistat, parthenolide, or alkyl MeLac) treatment of the cells was performed before the MeLac-alkyne probe treatment at a working probe concentration of 20 μM. Briefly, a pre-aliquoted storage stock solution of 20 mM, 50 mM, or 100 mM inhibitor (dissolved in DMSO, stored as aliquots at -80 °C) was diluted with DMSO to suitable concentrations to prepare working stock solutions if necessary. At each inhibitor concentration, an appropriate volume of each inhibitor working stock solution was then added to each prepared cell culture flask for each biological replicate (10 μL per T75 flask with 10 mL of PBS or 20 μL per T175 flask with 20 mL of PBS). Various inhibitor concentrations were investigated (1 μM vs. 10 μM for orlistat and parthenolide; 10 μM vs. 100 μM for alkyl MeLac). The inhibitor-treated cells were subject to similar quenching and washing steps, as described above.

3.2.4 Cell lysis and protein extraction

An appropriate volume of ice-cold PBS lysis buffer (6 to 10X of pellet volume) containing 1X EDTA-free protease inhibitor cocktail (Thermo Scientific) and phosphatase inhibitor cocktails (Thermo Scientific) was added to each cell pellet. The subsequent cell suspension was homogenized using an ultrasonic cell disruptor (VCX 130, Vibra-Cell) at 40% power on ice for three pulses of 5 seconds. The homogenized lysate in each conical tube was split and transferred

to multiple 2-mL microcentrifuge tubes for centrifugation at 16,000 g for 30 min at 4 °C. The lysate supernatants in microcentrifuge tubes were recombined for each original lysate sample, followed by BCA (Pierce) assay quantitation for total protein content. Finally, the cleared cell lysates were diluted with PBS lysis buffer to give a final protein concentration of 1.5 mg/mL for copper-catalyzed azide-alkyne cycloaddition (CuAAC) click reactions.

3.2.5 CuAAC click conjugation of probe-reacted proteins with azide tags

A click reagent cocktail was prepared *in situ* by mixing tris(benzyltriazolylmethyl)amine (TBTA, TCI) with a final concentration of 100 μM, copper(II) sulfate (Sigma-Aldrich, at a final concentration of 1 mM), TCEP (tris(2-carboxyethyl)phosphine, Pierce, at a final concentration of 1 mM), and the azide fluorescence or affinity tag of choice [at a final concentration of 25 μM; tetramethylrhodamine biotin azide (TAMRA), desthiobiotin azide (Des, **Scheme 3.5**), Dde biotin picolyl azide (Dde, **Scheme 3.6**), or diazo biotin azide (Dia, **Scheme 3.7**) from Click Chemistry Tools], in a tri-solvent system consisting of DMSO (Sigma-Aldrich), tert-butanol (Fisher Scientific), and water (MS-grade, Fisher Scientific). The click reagent cocktail was added to each sample of diluted probe-reacted lysate containing 1.5 mg/mL of protein. The reaction mixtures were incubated at room temperature for 1 hour with constant agitation. After the reaction, proteins were precipitated using cold acetone at -20 °C overnight. The resulting protein precipitate was pelleted by centrifugation, resuspended once in cold methanol using ultrasonic cell disruptor, re-pelleted, and washed with cold methanol two more times without sonication. The slightly air-dried pellet was then re-dissolved in 500 μL to 1mL of a reconstitution buffer (pH 8.0) containing 50 mM Tris-HCl (Fisher Scientific), 8 M urea (Pierce), and 10 mM 2-mercaptoethanol (Bio-Rad). Next, the protein samples were incubated with 10 mM dithioerythritol (DTE, ACROS Organics)

at room temperature for 45 minutes, followed by 18.8 mM iodoacetamide (G-Biosciences) at room temperature in the dark for 30 minutes. Finally, sodium dodecyl sulfate (SDS, Fisher Scientific) was added to each protein sample to a final concentration of 2%. Samples were directly used for either SDS-PAGE or affinity enrichment or stored at -80 °C.

3.2.6 Trypsin digestion of probe-reacted proteins and affinity enrichment of probe-modified analytes

For each probe-treated sample, affinity enrichment was performed at either the protein level (for the enrichment of probe-reacted proteins before trypsin digestion, performed for all three affinity tags) or peptide level (for the enrichment of probe-modified tryptic peptides after trypsin digestion, only performed for Des).

3.2.6.1 Affinity enrichment at the protein level

500 μL (containing about 0.75 mg of total protein) of each sample was transferred to a centrifuge column and diluted with 12 mL of 100 mM Tris-HCl buffer (diluting SDS to less than 0.1%, urea to less than 0.3 M). Next, 200 μL of pre-washed 50% NeutrAvidin-agarose resin (Pierce) slurry was added to each centrifuge column and incubated with constant agitation at 4 °C overnight. For each sample, the protein-bound resins were washed twice with 10 mL of washing buffer A (100 mM Tris-HCl, 0.1% SDS), three times with 10 mL of washing buffer B (100 mM Tris-HCl), and twice with 10 mL of LC-MS-grade water. Afterward, the probe-reacted proteins were eluted with one of three corresponding elution buffers prepared fresh. The Des elution buffer consisted of 0.1% SDS (or 4 M urea), 5 mM biotin (Sigma-Aldrich), and 20 mM TEAB; the Dia elution buffer consisted of 4 M urea, 37.5 mM sodium dithionite, and 20 mM TEAB; the Dde elution buffer consisted of 4 M urea, 3% hydrazine, and 20 mM triethylammonium bicarbonate

(TEAB, Sigma-Aldrich). The subsequent eluates were concentrated by a vacuum concentrator, extracted and digested using a tip-based solid-phase extraction/trypsin digestion protocol.

3.2.6.2 In-tip trypsin digestion of affinity enriched proteins

Solid-phase extraction/trypsin digestion (SxTd) tips were prepared in-house. To prepare each SxTd tip, two lays of Empore C18 material (two pre-stacked 3M Empore C18 Extraction Disks), and ten lays of glass fiber material (packed twice of five pre-stacked Whatman glass microfiber filters) into a 300 µL pipette tip (Fisherbrand, SureOne, Micropoint Pipette Tips, Universal Fit, Non-Filtered). After packing the SxTd tip, the sorbent was pre-wet twice with 200 µL of 0.1% trifluoroacetic acid in 80% acetonitrile, spun at 1000 xg, RT, for 2 min. The sorbent was then conditioned twice with 200 µL of freshly prepared 0.1M TEAB in 90% methanol, spun at 1000 xg, RT, for 2 min. The maximal loading capacity of each SxTd tip was 50 µL (containing up to 20 µg of total protein) affinity enrichment eluate sample.

For all protein samples with 5 to 20 µg of total protein in each, the total amount of needed trypsin was determined on the theoretical NeutrAvidin-agarose resin binding capacity (8 mg biotinylated protein per 1 mL of settled resin) and a trypsin:protein ratio (w/w) of 1:30. Fresh precipitation buffer (PrecB) was prepared by mixing methanol and 1 M TEAB at a ratio (v/v) of 9:1. Trypsin suspension was prepared by mixing 1.0 µg/µL trypsin stock with ice-cold PrecB at a volume ratio of 1:850. After brief vortexing, 280 µL of trypsin suspension was mixed with each 50 µL protein sample. Each suspension was then loaded to an SxTd tip, spun at 1000 xg for 3 min. The loaded SxTd tips were washed with 200 µL of ice-cold PrecB for three times, 200 µL of dry ice-cold acetone twice, and 200 µL of ice-cold PrecB once; all steps were followed by centrifugation at 1000 xg for 3 min. Afterward, 20 µL of freshly prepared ice-cold trypsin in 20 mM TEAB solution (containing 0.34 µg of trypsin) was added to each loaded SxTd tip to submerge

the sorbent fully (no air bubble). Finally, the loaded SxTd tips were placed in an appropriate pipette tip box with deionized water at the bottom (pipette tip not touching the water), sealed in a clean plastic bag, and incubated in a cell culture incubator at 37°C for 16 hrs. On the second day, each SxTd tip was quenched with 200 µL of 0.1% trifluoroacetic acid twice. The digest was eluted with 60 µL of 50% acetonitrile, spun at 1000 xg for 5 min, concentrated in SpeedVac (Thermo), lyophilized on a FreeZone Freeze Dry System (Labconco), and reconstituted in a solvent containing 0.1% formic acid (Fisher Scientific), 2.5% acetonitrile in LC-MS grade water for LC-MS/MS analyses.

3.2.6.3 Affinity enrichment at the peptide level

Buffer exchange was performed using Zeba spin desalting columns (7K MWCO, Pierce) according to the manufacture's instruction manual. After buffer exchange, the eluates (containing 50 mM ammonium bicarbonate, 2 M urea, and 1 mM calcium chloride, Sigma-Aldrich) from the spin columns were diluted to a resulting urea concentration of 0.8 M. The protein samples were then digested at a trypsin (sequencing grade, Pierce, or TPCK-treated, bioWORLD)-to-protein ratio of 1:50 or 1:25 and incubated at 37 °C for 16 hours. Next, pre-washed 50% NeutrAvidin-agarose resin slurry (100 µL per 0.75 mg of protein) was mixed with each digested sample, followed by 1-hour incubation at room temperature with constant agitation. The resins were then washed three times with 500 µL of Pierce IP Lysis Buffer, four times with 500 µL of PBS, and four times with 500µL of LC-MS-grade water. The captured peptides were eluted by the elution solvent (1.5X volume of the resin slurry, 50% acetonitrile, 0.1% trifluoroacetic acid, Fisher Scientific) with 3-minute incubation for three times. The subsequent eluate fractions were pooled in microcentrifuge tubes, concentrated in a SpeedVac, lyophilized, and reconstituted in a solvent

containing 0.1% formic acid (Fisher Scientific), 2.5% acetonitrile in LC-MS grade water for LC-MS/MS analyses.

3.2.7 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis and Western Blot

The gel-based analysis was performed on all the probe-reacted and avidin-enriched protein samples. A small portion of each sample (10 µL) was loaded onto a Mini-PROTEAN® TGX™ precast gel (1.0-mm, 10-well, Bio-Rad). Together with the samples, 5 µL of either SeeBlue™ Plus2 pre-stained protein standard (Life Technologies) or Precision Plus Protein™ dual color protein standards (Bio-Rad) was loaded to the gel as molecular weight references. The gels were prepared in duplicates for performing analytical electrophoresis and Western Blot in parallel. The subsequent electrophoresis was performed using Tris/Glycine/SDS premixed electrophoresis buffer (Bio-Rad) either at constant 120 V for 90 min until complete migration of the fronting dye band to the bottom of the gel. After electrophoresis, the protein bands were fixed with a solution of 10% acetic acid and 50% methanol in water. For the analytical electrophoresis, gels were stained with QC Colloidal Coomassie Stain (Bio-Rad), placed in a clear resealable plastic pouch, and imaged with a document scanner. For the Western Blot, the gel was removed from the electrophoresis cassette immediately after electrophoresis, rinsed with water, and accommodated in a sandwich assembly of pre-wet fiber pads, filter paper, PVDF membrane (Life Technologies; pre-soaked in methanol), and gel holder cassette. The sandwich assembly was then placed together with a pre-frozen cooling pack and a magnetic stirrer in the buffer tank containing freshly prepared transfer buffer (10% methanol in Tris/Glycine/SDS premixed electrophoresis buffer). The transfer was performed at constant 30V overnight. After completion of the transfer, the PVDF membrane was removed from the sandwich assembly and placed in a clean square petri dish provided by the

WesternBreeze kit (Invitrogen). Solutions for chromogenic immunodetection were prepared according to the WesternBreeze instruction manual. Briefly, the membrane was incubated on an orbital shaker (Thermo Fisher Scientific) at 1 rev/s with: *1)* 20 mL of water for 5 min; *2)* 10 mL of blocking solution for 30 min; *3)* 20 mL of water for 5 min (twice); *4)* 10 mL of primary antibody (diluted to 1 mg/mL from original 8.17 mg/mL solution, mouse monoclonal anti-biotin IgG, Fitzgerald) solution (0.05 µg/mL, 1:20000) for 1 h; *5)* 20 mL of antibody wash solution for 5 min (4 times); *6)* 10 mL of secondary antibody- alkaline phosphatase conjugate (anti-mouse IgG, Life Technologies) for 30 min; *7)* 20 mL of antibody wash solution for 5 min (4 times); *8)* 20 mL of water for 2 min (3 times); *9)* 5 mL of chromogenic substrate solution for 1 h; and *10)* 20 mL of water for 2 min (3 times). Finally, the blotted PVDF membrane was dried in air and scanned with a document scanner.

3.2.8 Liquid chromatography-tandem mass spectrometry methods

For each sample**,** enriched and reconstituted tryptic digests were analyzed on a Q Exactive Plus mass spectrometer equipped with the nanospray ionization (EASY-Spray nano) source coupled with an EASY-nLC 1200 system (Thermo Fisher Scientific). For the LC part, peptides were separated on a fused silica capillary (30 cm x 100 um I.D) packed with Halo C18 (2.7 um particle size, 90 nm pore size, Michrom Bioresources). The autosampler temperature was 4.0 °C. Column oven temperature was 50.0 °C. Sample injection volume was 2.0 µL. Mobile phase flow rate was set as 300 nL/min. Solvent A was 0.1 % formic acid in water, and solvent B was 0.1 % formic acid in acetonitrile. A gradient of 0 to 40% solvent B over 180 min was applied. For the mass spectrometer part, the Xcalibur v2.8 software was used to control the instrument. The mass spectrometer operated in DDA mode for monitoring positive ions at a spray voltage of 2200 V. For MS1, the mass range was from 350 to 1700 *m/z*. The Orbitrap mass analyzer was set with a

resolution of 70,000, an AGC target of 1E6, and a maximum ion time of 100 ms. For data dependent MS2, the quadrupole was set with an isolation window of 2.0 *m/z*. The Orbitrap was set with a resolution of 17,500, an AGC target of 1E5, and a maximum ion time of 50 ms. This DDA method allowed up to 10 MS/MS scans per duty cycle, and a stepped normalized collision energy (NCE) of 27. Precursors that triggered MS/MS scans were dynamically excluded from repetitive MS/MS scans for 40 s. Charge state exclusion was enabled to reject precursor ions with charge states outside the range of +2 to +4. The peptide match option was set as preferred. MS/MS spectra were collected as the profile data type.

3.2.9 Mass spectrometry data qualification and analysis

For peptide and protein identification, the data was initially searched against a canonical human reference proteome database (a FASTA file, Swiss-Prot, Homo Sapiens, UP000005640, last modified on January 26th, 2019) containing 20,425 protein sequences. DDA raw files were analyzed with both FragPipe[224] (Version 12.2, with MSFragger Version 2.3) for unrestricted database search (open-search) and MaxQuant[225] (Version 1.6.5.0) for restricted database search with distinct settings. For more comprehensive identification and quantitation of probe modification site-specific peptides, the same MaxQuant database search was also performed with an extended proteome database (a FASTA file, UniProt, Homo Sapiens, UP000005640, last modified on May 7th, 2020) containing 210,599 protein sequences compiled of both canonical and isoform protein entries.

3.2.9.1 Unrestricted database search for a global illustration of MeLac and MeLac-related chemical modifications

The open search of processed data was performed on all the samples using MSFragger. The raw DDA data files were first centroided and converted to mzXML open-source format using

ProteoWizard[167]. To perform the search, a set of default open-search parameters was then used except for the items as follows: a precursor *m/z* tolerance (DeltaMass) window of 0 to 1000 Da, a peptide length range of 5 to 50 residues, a peptide mass range of 500 to 5000, a maximum missed peptide cleavage of 1. Both variable and fixed modification inclusion lists were disabled. Finally, mass shifts (of precursor ions, caused by chemical modifications in samples) in the open-search results were visualized using the DeltaMass[226] software tool.

3.2.9.2 Restricted database search for identification of MeLac-reacted proteins

For the identification of target proteins for MeLac-alkyne probe **1**, the MaxQuant search was performed on probe-treated samples that underwent the protein-level enrichment procedure. The corresponding search parameters were less than 1% peptide-level false discovery rate (FDR), less than 1% protein-level FDR, less than 1% modification site FDR, a minimum peptide length of 5, a minimum score of 0 for unmodified peptides, a minimum score of 40 for modified peptides, a minimum unique peptide number of 2, a minimum razor (+ unique) peptide number of 2, a minimum peptide number of 2, an MS/MS mass error tolerance of 20 ppm, a peptide length range of 8 to 25 for unspecific search, a maximum missed peptide cleavage of 1, a maximum peptide mass of 4600 Da, and a revert decoy mode. A basic modification inclusion list (variable oxidation on methionine, variable acetylation on protein N-terminus, and fixed carbamidomethylation on cysteine) was used with up to 3 modifications per peptide.

3.2.9.3 Restricted database search for identification of MeLac-modified peptides

For the identification of MeLac-alkyne probe-modified peptides and their modification sites, the MaxQuant search was performed on probe-treated samples that underwent peptide-level enrichment procedure. To limit the search space, a separate search was performed on each dataset of these samples for each expected reactive amino acid residue (cysteine, lysine, serine, threonine,

or tyrosine, respectively). Apart from the common settings reported above, the distinct search parameters were a minimum unique peptide number of 0, a minimum razor (+ unique) peptide number of 1, a minimum peptide number of 1. The modification inclusion list was set according to the expected residue to be modified. For cysteine-specific search, a customized modification inclusion list covering variable oxidation on methionine, variable acetylation on protein N-terminus, variable carbamidomethylation on cysteine, variable original Des-MeLac (+550.3115) on cysteine, and variable hydrolyzed Des-MeLac (+568.3221) on cysteine was used with up to 5 modifications per peptide. For non-cysteine-specific search, a customized modification inclusion list was used with up to 5 modifications per peptide, covering variable oxidation on methionine, variable acetylation on protein N-terminus, fixed carbamidomethylation on cysteine, variable original Des-MeLac (+550.3115 on either lysine, serine, threonine, or tyrosine), and variable 2-mercaptoethanol-quenched Des-MeLac (+628.3254 on either lysine, serine, threonine, or tyrosine). Two reported desthiobiotin-PEG_3-specific fragment ions were set as diagnostic peaks (197.1285 Da and 240.1707 Da).[227,228] Example spectra were visualized and modified for formal presentation using IPSA.[229]

3.2.9.4 Identification of glutathione Des-MeLac-modified peptides

For the identification of glutathione Des-MeLac-modified peptides and their modification sites, the MaxQuant search was performed on probe-treated samples that underwent peptide-level enrichment procedure. A separate custom human proteome database (a FASTA file, Swiss-Prot, Homo Sapiens, UP000005640, last modified on April 9th, 2020, appended with GSTP1-I105 → V105) containing 20,351 protein sequence entries was used (before the switching to the extended UniProt database) as the reference proteome for covering the isoform-specific peptide YVSLIYTNYEAGK, which was discovered by the manual examination of previous database

search results. Reactive amino acid residues: cysteine, lysine, serine, threonine, and tyrosine were included. Other settings were a minimum unique peptide number of 0, a minimum razor (+ unique) peptide number of 1, a minimum peptide number of 1. For the customized modification inclusion list covering variable oxidation on methionine, variable acetylation on protein N-terminus, variable carbamidomethylation on cysteine, and variable glutathione Des-MeLac (+857.3953) on cysteine with up to 5 modifications per peptide. Two reported desthiobiotin-PEG_3-specific fragment ions were set as diagnostic peaks (197.1285 Da and 240.1707 Da).[227,228]

3.2.9.5 Label-free quantitation of probe-modified peptides and visualization of orlistat, parthenolide and alkyl MeLac competitive ABPP data

The relative abundance of each identified/sequenced tryptic peptide was quantified and calculated as the extracted chromatogram peak area of its corresponding precursor at the MS1 level as part of the MaxQuant database search workflow. The MaxQuant search output was then processed and summarized as tables of fold change vs. p-value by the in-house developed R package maxabpp. The data was finally visualized as volcano plots.

3.2.10 Quantum mechanical modeling

All computational studies were performed using Gaussian 09 (Revision D.01). All the input model structures were built with GaussView 5.0.9.

3.2.10.1 MeLac reaction paths

MeLac reaction paths were modeled with three different protein nucleophiles: thiol, hydroxyl, and amino groups. Simplified model structures, methyl MeLac, methanethiol, methanol, and methylamine, as well as their corresponding six transition states (TS) and six products, were built for either Michael addition or acyl addition reaction path. Geometric optimization of these structures was then performed with a range-separated DFT functional ωB97XD, employing a

correlation-consistent Dunning basis set aug-cc-pVTZ for effective modeling of Michael addition reaction path for the thiol.[230] The Self-Consistent Reaction Field (SCRF) used Polarizable Continuum Model (PCM) as an implicit method to account for the polarizable solvent effect of water on reactions. Afterward, intrinsic reaction coordinate scan (IRC) computation was performed on each TS structure to ensure its energy saddle point location on the reaction path. Finally, vibrational frequency computation was performed on all optimized structures to determine thermal corrections to their Gibbs free energy.

3.2.10.2 Des signature ions

The geometry minimization of Des signature ions was performed with a basic DFT functional B3LYP, employing a Pople basis set 6-311+g(d, p) of moderate size. The output structures of the lowest total energy (f_1 and f_2) were consistent with the previously reported fragments.[228]

3.2.11 Computational protein-ligand modeling

The computational protein-ligand modeling (molecular docking) experiment was performed using Schrödinger Suite (2020-1 Release).

3.2.11.1 Ligand and protein preparation

Four GSH-Lac β-lactone stereoisomers, *3R,4R*-, *3R,4S*-, *3S,4R*-, and *3S,4S*- were generated on ChemDraw (version 19.0) as two-dimensional (2D) structures with specified chirality on both the β-lactone and the GSH moieties. Their corresponding SDF files were imported to the Maestro (version 12.3) workspace as ligands. Next, three-dimensional (3D) coordinates of all ligands were generated by the LigPrep module with protonation states dominant within a pH range of 7.0 ± 2.0, preserving their tautomeric forms. The LigPrep output then served as starting ligand geometries for docking.

The crystal structure of a human glutathione S-transferase P dimer (GSTP1) complexing two hydrolyzed piperlongumine ligands (PDB ID: 5J41)[231] obtained from Protein Data Bank (PDB) was subjected to several preparation steps on the Protein Preparation Wizard in Maestro before used as protein model for docking: removal of surface water molecules, assignment of bond orders, and addition of missing hydrogen atoms. Next, the orientation of amide (Asn and Gln), hydroxyl (Ser, Thr, and Tyr), and thiol groups (Cys) and the protonation and tautomeric state of His residues on the protein model were also optimized. In the final preparation step, restrained minimization of the protein model was applied using 0.3 Å RMSD constraint and OPLS3e force field. The Protein Preparation Wizard output then served as the starting protein geometry for receptor grid generation.

3.2.11.2 Receptor grid generation

For the starting protein geometry, a ligand-enclosing box was centered on one of the piperlongumine ligands so that the ligand to be docked (GSH-Lac) would be confined to the enclosing box. The size of the enclosing box was set to be dependent on the piperlongumine ligand.

3.2.11.3 Ligand docking

Docking calculations involved Glide Extra Precision (XP) with default sampling, pre-processing, and postprocessing. The preprocessing refers to the restrained minimization of the model GSTP1-piperlongumine complex in the OPLS3e force field, which is part of the protein preparation process. Postprocessing refers to post docking minimization of the GSTP1-GSH-Lac complex. Glide was set to write the 5 best poses per ligand.

3.3 Results and Discussion

3.3.1 Broad reactivity of MeLac delivering wide proteome coverage

The broad reactivity of MeLac was characterized at four levels of resolution according to the resolving power of each specific analytical method and scope of data interpretation applicable throughout the workflow of this study. These resolution levels were the gel-based proteome resolution, MS-based protein resolution, MS-based peptide resolution, and MS-based amino acid residue resolution. First, the fluorescent TAMRA tag provided a gel-based shortcut to reveal the in-proteome reactivity of MeLac. Upon CuAAC conjugation to MeLac-alkyne probe-reacted proteins, the TAMRA fluorophore enabled direct detection of those proteins after gel electrophoresis. As a result, the fluorescence gel image of *in vitro* probe-treated cellular proteome (**Figure 3.2A**) displayed multiple fluorescent protein bands throughout gel lanes. This gel image suggested the successful labeling of various proteins by the MeLac-alkyne probe. The labeled proteins had a wide range of molecular weight from 250 kDa to 10 kDa, implying a wide proteome coverage of the MeLac-alkyne probe and broad reactivity of the MeLac warhead. Full lanes of protein bands were observed for proteome samples treated by the MeLac-alkyne probe at a probe concentration as low as 0.5 μM, indicating the in-cell applicability of the probe. The extent of background probe-protein reactions, where the probe labeled surface residues instead of ligand-binding/active site residues on proteins, was unclear. However, by comparing the rate of darkness diminishing of individual protein bands over the decline of probe concentration, the probe did potentially offer some selectivity towards certain proteins with molecular weight around 50 kDa.

3.3.2 Interpreting MS-based profiling data: probe-protein reactions characterized at multiple levels of resolution

Over the gel-based analysis, MS-based proteomics profiling provides more details of probe-protein reactions. In chemical proteomics, complex protein samples are preferentially

analyzed as tryptic digests via the bottom-up approach by HR/AM mass spectrometers. As part of the affinity capture methodology and sample processing workflow, biotin-avidin-based affinity enrichment techniques are indispensable. These affinity enrichment techniques can significantly improve the detectability of low-abundance analytes. In practice, the enrichment of analytes can be performed either before the trypsin digestion of protein samples (at the protein level) or after trypsin digestion of those samples (at the peptide level). If captured at the protein level, the enriched analytes are mapped to a comprehensive picture of the whole target protein candidate pool of the chemical probe. For each target candidate in the pool, multiple tryptic peptides are measured both qualitatively and quantitatively. According to the sequence coverage and number of detected unique peptides, indistinguishable proteins (usually isoforms or homologous proteins within the same family) are aggregated as a single identification hit. Once defined by the proteomics database search engine, these aggregated identification hits are usually treated as indivisible protein groups for downstream qualitative and quantitative analyses. Therefore, the identification of individual proteins is more straightforward and less proteins are grouped if the sequence coverage is high and number of sequenced unique peptides is large for each protein. On the flip side, the protein-level enrichment process almost always produces less-than-ideal preparations adulterated with non-probe target background proteins. The double aggravated sample complexity, by both tryptic peptides from background proteins and unmodified peptides from probe-reacted proteins, usually fails the detection of probe-modified peptides. In contrast to protein-level enrichment, the peptide-level enrichment process concentrates probe-modified peptides further at the cost of sequence coverage and uniqueness enforcement of individual target proteins.

For this study, the MS-based characterization of protein reactions with the MeLac-alkyne probe is defined at three levels of resolution according to the assignability of reaction sites on specific proteins, residues, or peptides. Complete characterization of a protein reaction with a covalent chemical probe ultimately requires localizing the probe modification on a specific amino acid residue of a protein (residue resolution). Both the protein-level and peptide-level affinity enrichment were performed on MeLac-alkyne treated samples. When the affinity enrichment was performed at the protein level, few probe-modified peptides were detected. Consequently, the assignment of probe-protein reactions could solely depend on the identification of reacted proteins (protein resolution) in enriched samples. There are two reasons. Some probe-reacted proteins may not produce probe-modified peptides that are suitable for bottom-up proteomics, the method-of-choice for chemical proteomics. Not all precursor ions of suitable peptides are sampled for MS/MS analysis in proteomics profiling. When the affinity enrichment was extended to the peptide level, for a few target proteins, it remained challenging to localize their probe modification sites because of several issues. A probe-modified peptide, in comparison to its native counterpart, does not always produce adequate sequence ions during its gas-phase fragmentation within the mass spectrometer. For each peptide precursor ion acquired for gas-phase fragmentation and MS/MS scans, the whole set of these sequence ions is essential evidence to localizing the probe reaction site at a specific residue. When the observable set of sequence ions is truncated as the probe modification perturbs the gas-phase fragmentation behavior[232] of probe-modified peptide, the site localization can be ambiguous. In this case, it is plausible to only assign the probe reaction site to the modified peptide (or its fragment) instead of a specific residue on the corresponding protein. Consequently, the corresponding protein reaction is characterized at a moderately reduced resolution of the peptide level.

Depending on chemical properties of the chemical probe and affinity enrichment method of choice in various chemical proteomics studies, not all probe-protein reactions can be characterized based on probe-modified peptides. In some cases, probe-protein reactions can only be characterized at a heavily reduced resolution of the protein level. For instance, when a labile covalent probe is used in compound-centric chemical proteomics, target proteins are universally characterized at the resolution of the protein level. With decreasing resolution from the residue to protein level, tightened constraints must be set to determine confidence in reporting reactivity of the probe, which demands new analytical methods to evaluate confidence in the identification of the probe-reacted proteins.

3.3.3 Affinity tagging triplication differentiating target candidates at the protein level

To confront the analytical challenge in identifying and reporting probe-reacted proteins at the protein-level resolution, a novel method named affinity tagging triplication was developed to differentiate MeLac-alkyne probe-reacted proteins with discrete analytical confidence, even without detection of any probe-modified peptide. In each of three separate preparations (step 3, **Scheme 3.4**), a different (desthio)biotin tag was used for avidin-based affinity enrichment of probe-reacted proteins from one lysate sample of MeLac-alkyne probe-treated HT-29 cells. The rationale was that the tagged proteins underwent different releasing conditions to recover MeLac-alkyne probe-reacted proteins. Consequently, the resulting sample matrices and background proteins in the preparations obtained would be significantly different. When a protein was identified in more than one enrichment preparation, the confidence of it being a true MeLac-alkyne probe-reacted protein was increased. Specifically, the MeLac-alkyne probe-reacted proteins were tagged, using CuAAC click chemistry, with desthiobiotin azide (Des, **Scheme 3.5**), Dde biotin

picolyl azide (Dde, **Scheme 3.6**) azide or diazo biotin azide (Dia, **Scheme 3.7**). Tagged proteins were captured by NeutrAvidin-agarose resins and then released using different elution conditions for preparing trypsin digests.

Subsequently, a unique three-tier system was set based on the number of reproduced identifications of a particular protein in the differentially tagged preparations from a common lysate sample. The lysate was prepared from HT-29 cells upon in-live-cell reactions with MeLac-alkyne. In total, a protein target candidate pool of 1,505 canonical proteins (distinguished by gene names)—related to 2,074 protein groups (distinguished by aggregated UniProt accession numbers)—was identified by proteomics profiling and MaxQuant[225] database search (**Table ia, ib, and ic, Appendix**) at a threshold of 1% maximal peptide/protein FDR and two minimal unique peptides. The identified proteins were categorized in Tier I, II, or III (**Figure 3.2B** and **3.2C**) with decreasing confidence in being true probe-reacted proteins, accordingly. Proteins in Tier I were identified in all the three tagged samples; Tier II in two; Tier III for only one. Analogous to the minimum requirement of two unique peptides for identifying a precursor protein within a proteome, Tier I and II proteins are reported as confident identifications; Tier III proteins need further investigation. Notably, the three-tier categorization of identified proteins was performed with either gene names or protein groups of aggregated UniProt accession numbers. When categorized with gene names which did not annotate protein isoforms (**Figure 3.2B**), 17.5% proteins (264 out of 1,505) were identified as Tier I proteins, 25.1% proteins (378 out of 1,505) were identified as Tier II proteins, and 57.3% proteins (863 out of 1,505) were identified as Tier III proteins. When categorized with UniProt protein accession numbers, which annotated protein isoforms (**Figure 3.2C**), isoform-specific proteins in each protein group were compared across confidence tiers using maxabpp. In contrast to the convention where protein IDs are treated as a

series of characters and compared for exact matches, the maxabpp matching algorithm treats aggregated protein IDs as a group of individual entries. When comparing two protein groups, if maxabpp detects the existence of one or more shared individual protein(s), it will mark these two protein groups as related. Consequently, if unequal numbers of aggregated protein isoforms are identified in all three differentially tagged samples of unique sample matrices, these related protein isoforms will be categorized as Tier I proteins. With the implementation of this matching algorithm, higher numbers of intersecting proteins were observed as identified isoforms and homologous sequences. 34.3% proteins (711 out of 2,074) were identified as Tier I proteins, 31.5% proteins (653 out of 2,074) were identified as Tier II proteins, and 34.2% proteins (710 out of 2,074) were identified as Tier III proteins (**Figure 3.2C** and **Table ii, Appendix**). Overall, proteomics profiling of enriched MeLac-alkyne probe-reacted proteins supported the broad reactivity of the MeLac warhead with proteins in the proteome, as observed in gel-based ABPP (**Figure 3.2A**), but offered limited information for understanding protein reactions with MeLac.

3.3.4 Reactivity investigation to deepen at peptide the level

Proteomics profiling of enriched MeLac-alkyne probe-modified peptides confirmed the broad reactivity of MeLac. A total of 751 MeLac-alkyne probe-modified peptides with 1,430 modification sites were identified by MaxQuant[225], resulting in the identification of 562 canonical proteins (**Figure 3.2D** and **3.2E**; **Table iii, Appendix**). The Andromeda[233] peptide-spectrum match (PSM) scoring algorithm of MaxQuant was set to filter PSM hits with a minimum score cutoff of 40 and 1% false discovery rate threshold for both peptide and protein identification. Consequently, for most peptides, alternative PSM assignments with lower scores were discarded. Among all 1,430 MeLac modification sites, a fraction of sites had ambiguity in localization on peptides; reactions could only be assigned to fragments of modified tryptic peptides.

Approximately 1,000 peptides had localized modification sites. Moreover, MeLac-alkyne probe-modified peptides originated from functionally diverse proteins, covering enzymatic reactions of all seven general types; these proteins were modified at either catalytic or non-catalytic residues (**Figure 3.3**, **Figure 3.4**, and **Table 3.2**). Interestingly, in comparison with 1,505 gene name-annotated canonical proteins identified in samples from protein-level enrichment experiments, a significant number (292 out of 1,505) of proteins were identified only from samples of enriched MeLac-alkyne probe-modified peptides. Apart from the existence of background peptides, this observation implied that the detection of low-abundance peptides, upon probe modification, required further concentration at the peptide level. The affinity pull-down of these peptides successfully reduced sample complexity and achieved higher identification rate of probe-modified peptides.

Identification of MeLac-alkyne probe-modified peptides had its own challenge in confidence, differing from the confidence challenge for enriched samples of MeLac-alkyne probe-reacted proteins. The identification of the modified peptides depended only on a single amino acid sequence. Higher-energy collisional dissociation (HCD) of peptide ions used in the mass spectrometer of this work generated incomplete sequence ions, preventing *de novo* sequencing of the peptides. Therefore, additional constraints should be imposed for verifying MS/MS spectra and identifying MeLac-alkyne probe-modified peptides. These constraints were particularly important because a relatively low PSM score of 40 was set for the database search and identification of modified peptides. The verification of a spectral match would require the presence of two signature ions from Des as the modification-specific fragment ions (labeled on spectra as f_1 and f_2, **Figure 3.2F** and **Scheme 3.8**). Specificity of these two fragment ions was high (**Table 3.3**), enabling verification of identified MeLac-alkyne probe-modified peptides because Des was

introduced to the modified peptide as a xenobiotic moiety via CuAAC attachment of the Des tag for enriching modified peptides. However, not all identified spectra of probe-modified peptides had a lower mass cutoff to detect the signature ions, due to the dynamic mass range setup during MS analysis (white vs. red areas in **Figure 3.2E**). The gene name-annotated canonical proteins identified based on probe-modified peptides (red slides in **Figure 3.2D**) were compared to the identified ones from protein-level enrichment samples (**Figure 3.2B**). A decreasing trend on the percentage of identifiable probe-reacted proteins from Tier I to III (30.6% to 21.4% to 12.5%, respectively) supported the effectiveness of the affinity tagging triplication strategy for identifying true MeLac targets. Although the Tier III proteins should be treated with lower confidence, observing probe-modified peptides provided a means to separate the protein targets from the remaining pool of proteins, which had the largest fraction of background proteins. Proteomics profiling of enriched peptide samples also identified significant numbers of proteins in each tier (orange slides, **Figure 3.2D**) based on peptides, which did not carry MeLac modifications in the samples; 36.4% for Tier I, 21.2% for Tier II and 18.4% for Tier III, respectively. The decreasing percentages were consistent with the declining confidence levels of Tier I, II, and III proteins being true MeLac targets. Overlapping sequences between probe-modified and non-probe-carrying peptides indicated a possible loss of the MeLac modification during sample preparation. For instance, acyl addition by nucleophilic thiol produces labile thioesters.[220] Non-specific peptides carried through the peptide enrichment preparation could also contribute to the observed non-probe-carrying peptides. However, the fact that more than one third (inner orange slide, **Figure 3.2D**) of the Tier I proteins were also identifiable based on the non-probe-carrying peptides suggested that the non-specific peptides could be a minor concern.

MS/MS spectra of MeLac-alkyne probe-modified peptides provided direct and detailed information for characterizing probe-protein reactions. MS/MS spectra with unambiguously localized sites for modification were used to explain the MeLac chemistry. Modification of peptides by MeLac-alkyne was attributable to both Michael addition and acyl addition of MeLac by Cys, Ser, Thr, Tyr, and Lys (**Scheme 3.9 and Appendix**). Despite being uncommon and thermodynamically unfavorable based on computation results (**Figure 3.5**), Michael addition modifications of Lys and Thr have been reported for proteins.[234,235] Possible nucleophilic substitutions of the lactone ring of MeLac (mechanism 2b on **Scheme 3.2**), whose reaction products would have degenerate masses as those from acyl addition, were not considered in this work. Upon chemical modification by MeLac-alkyne and downstream CuAAC tagging with Des, the modified peptides had three distinct mass shift values. A global view of MeLac modifications on peptides was apparent on a DeltaMass plot based on unrestricted MSFragger search[226] (**Figure 3.6A**). Upon CuAAC attachment of the Des tag, the modification of a residue by MeLac-alkyne led to a total mass shift of 550.3115 Da (Des MeLac, top of **Figure 3.6B**). Depending on the reaction site on MeLac, further modifications by Des MeLac on reacted peptides occurred during sample preparation: hydrolysis of the lactone ring causing an additional increase of 18.0106 Da (568.3221 Da in total, bottom of **Figure 3.6B**) and quenching of the Michael receptor with 2-mercaptoethanol adding a further increase of 78.0139 Da (628.3254 Da in total) as in **Figure 3.6C**. Modified peptides with a mass increase of 550.3115 Da were attributed to the products of Michael addition (**Scheme 3.9**), but not acyl addition; an excessive amount of 2-mercaptoethanol was used for quenching the Michael acceptor and the thioester bond formed via acyl addition of the β-lactone ring is labile.[220] Indeed, few modified peptides could be assigned as quenched adducts from the acyl addition of Des MeLac by Cys.

3.3.5 Advantages of MeLac-alkyne in probing reactive cysteine

Cys residues had the largest number of MeLac modification sites (**Figure 3.2E**). The cysteinome provides many important targets for covalent drugs.[119] The majority of MeLac-alkyne probe-modified peptides carried a localized site of Cys modification. This observation was not a surprising fact considering the large numbers of reactive cysteines under physiological conditions. The profiling results of MeLac-alkyne were compared with those of iodoacetamide (IA)-alkyne, a routinely used broad-spectrum cysteine-reactive probe in chemical proteomics. Localized IA reaction sites on 6143 cysteinyl peptides had been reported, among which 758 sites were ligandable.[77] Although sharing 128 ligandable peptides with the IA-alkyne pool, 224 out of the 653 peptides with probe modification, either catalytic or non-catalytic cysteines, were unique to MeLac (**Figure 3.7**). It is important to note that besides the chemical difference between MeLac and IA, MDA-MB-231 and Ramos cells were used in the IA study. Although known as highly cytotoxic and incompatible with live cells, the IA-alkyne probe was used at a concentration of 100 μM to treat live cells.[77] In comparison, HT-29 cells were treated with MeLac-alkyne at concentrations up to 50 μM for gel-based analysis and 20 μM for MS-based analysis in this work.

A significant advantage of MeLac over IA probes comes from differences in cytotoxicity. IA has high cytotoxicity,[236] limiting the application of IA probes mainly to lysates.[237] In contrast, MeLac compounds were used with live cells at concentrations up to 100 μM for MeLac-alkyne and 300 μM for alkyl MeLac inhibitor in this work. Cells remained adhered to culture flask surface under these concentrations for incubation times up to 1 hour. Two important analytical advantages from the live-cell application of a measurement probe are (1) decreased background reactions and (2) enabled analysis of proteases whose activity must be blocked during cell lysis. Additionally,

MeLac-based competitive ABPP for analyzing the action of non-covalent inhibitors in live cells is a unique potential.

3.3.6 Establishing a versatile competitive ABPP platform using MeLac-alkyne probe and peptide-centric quantitation approach

This study then proceeded to test the major designed utility of MeLac probes in competitive ABPP platforms for analyzing proteome-wide reactions of reactive molecules. Three inhibitors: orlistat, parthenolide, and alkyl MeLac inhibitor, were analyzed the competitive ABPP platform using the MeLac-alkyne probe. The orlistat and parthenolide were selected as model covalent inhibitors, while the alkyl MeLac was introduced as a MeLac parent compound or putative MeLac-based inhibitor for studying the chemoselectivity of the MeLac warhead.

Orlistat is a β-lactone compound capable of forming covalent adducts with nucleophilic hydroxyl and thiol groups on proteins. It is an FDA-approved drug that targets lipases for weight management. Extensive chemical proteomics analysis has revealed many off-target proteins,[73,220], which are linked to anti-tumor activities[238] and organ toxicity[239] for this over-the-counter drug. Most of the covalent orlistat-protein adducts are formed via a labile thioester bond.[220] Confident identification of off-target proteins of orlistat, especially orlistat-modified peptide, is extremely challenging due to the fact that the thioester or ester bond is not stable during sample preparation for proteomics analysis. This is a common problem for β-lactone based probes. Live-cell competitive ABPP using MeLac-alkyne as the measurement probe provides a valuable solution to this analytical challenge.

Parthenolide is an α-methylene-γ-lactone natural product with anti-inflammation and anti-tumor activity.[240-243] Parthenolide has a Michael acceptor subject to nucleophilic addition as well; both exist in the MeLac warhead. Unlike β-lactone, which is prone to nucleophilic attacks and

ring-opening reactions, γ-lactone is significantly more stable and inert.[244] Although a systematic chemistry study comparing the reactivity of the Michael acceptor and γ-lactone of parthenolide is unavailable, the existing cysteine-dominant parthenolide reactivity profile[240] suggests that parthenolide would only react with protein nucleophiles as a Michael acceptor for nucleophilic addition reactions.

Proteome-wide inhibitor selectivity and protein dose responses were first explored with gel-based competitive ABPP at various inhibitor concentrations over a wide concentration window. Subsequently, two reasonable concentrations were selected based on the gel-based ABPP results (**Figure 3.8**) for MS-based quantitative experiments for each inhibitor. Orlistat, parthenolide (1 vs. 10 μM), or alkyl MeLac (10 vs. 100 μM) treatment, live HT-29 cells were incubated with MeLac-alkyne at 20 μM. The probe-modified peptides were then enriched for MS-based profiling. In total, 395 probe-modified peptides from 316 proteins were quantified for samples from the orlistat competition experiment (**Figure 3.9A**); 320 probe-modified peptides from 276 proteins were quantified for parthenolide (**Figure 3.9B**); 44 probe-modified peptides from 41 proteins were quantified for alkyl MeLac (**Figure 3.9C**). Pairwise quantitative comparison of enriched probe-modified peptide samples from cells pre-treated with either orlistat, parthenolide at 10 μM vs. 1 μM, or alkyl MeLac at 100 μM vs. 10 μM was performed using a method of label-free quantitation (LFQ). For each quantified probe-modified peptide, the relative amount was determined as the mean of normalized areas of MS1 chromatographic peaks of precursor ions for the peptide in six replicates (two biological replicates with three analytical replicates each). The normalization was performed against the total ion intensity of all quantified ions of probe-modified peptides to correct run-to-run variations. In addition, if a longer miscleaved sequence was observed for a particular tryptic peptide with the probe modification, the peak area for this peptide was

calculated as the sum of the miscleaved and fully cleaved. Volcano plots of Log_{10} p-value vs. Log_2 FC were based on quantified probe-modified peptides (**Figure 3.9**). An FC cutoff threshed was not set for selecting protein targets of orlistat, parthenolide, or alkyl MeLac due to two reasons. The quantitative comparison was only performed within a tentative dose window, 10 μM vs. 1 μM, or 100 μM vs. 10 μM, which could be at any location on a dose-response curve for the inhibitory potency of the two inhibitor molecules. LFQ is less precise than isotope-labeling based quantitation by design. At the 5% statistical level of significance (p-value $\leq$ 0.05), 40 peptides/proteins responded to the increased orlistat concentration, and 16 peptides/proteins responded to the increased parthenolide concentration. Intriguingly, the parthenolide target candidate pool highlighted six previously reported protein targets, DCTN4, RPL4, RPL14, RPL18, RPS2, and PRKDC, (high-lighted in red, **Figure 3.9B**) from a label-based competitive ABPP experiment using a different human cell line and IA-alkyne probe.[240] Remarkably, the DNA-dependent protein kinase catalytic subunit (PRKDC) was reported as one of three major kinase targets showing high dose response (with a log_2 FC value of -5) to parthenolide.

In contrast to orlistat and parthenolide, the alkyl MeLac competition experiment had the smallest number of quantified proteins. This could be explained with: (1) Cells were treated with alkyl MeLac at concentrations 10-fold higher than the other two inhibitors, resulting in a less informative dose window capturing the region of dose saturation on the dose response curve; (2) alkyl MeLac shared the same reactivity with MeLac-alkyne, thus completely labeled and occupied all possible sites on protein targets of MeLac-alkyne at a higher concentration, resulting in detection of few MeLac-alkyne probe-modified peptides. In addition, probe-modified peptides were not detected for a group of proteins at the higher inhibitor concentration, particularly for alkyl MeLac, where 15 out of 16 statistically significant proteins were not detected at 100 μM. These

proteins formed an off-chart list of protein targets of the inhibitor used in the competition experiment. It was highly likely that these proteins were the best targets for the corresponding inhibitor. However, this statement would require further experimental verification considering the stochastic sampling of peptide ions during MS analysis. Notably, the selectivity profiles for orlistat, parthenolide, and alkyl MeLac differed significantly, exemplifying that the broadly reactive MeLac-alkyne probe assembles a versatile peptide-centric platform of competitive ABPP capable of analyzing reactive molecules with different chemical properties. The capability of MeLac reacting with multiple nucleophilic residues and measuring enzymes of all classes (colored dots in **Figure 3.9**) endorsed its foreseeable versatility.

3.3.7 MeLac warhead recruiting glutathione in live cells to assemble a selective β-lactone probe

Intriguingly, multiple electrophilic sites on the MeLac scaffold made it possible for MeLac-alkyne to conjugate a reactive affinity ligand on one site and use another site to react with protein nucleophiles. Considering the diversity of intracellular reactive metabolites, the unrestricted database search was initially used to screen for unknown protein modifications. Surprisingly, the MeLac-alkyne probe recruited endogenous glutathione to form a β-lactone probe that had much increased selectivity as the DeltaMass plot highlighted an unexpected modification with an unusual mass shift of 857 Da (**Figure 3.10A**). The high mass region of the DeltaMass plot also showed several other mass shifts with noticeable peptide spectra matching densities. The mass shift of 857 Da, being the most significant one, was attributed to the direct addition of GSH presumably at the Michael acceptor on MeLac to form a new β-lactone probe, namely GSH-Lac in live cells (**Scheme 3.10**). A dedicated MaxQuant search (with 857 Da as the input modification parameter) identified 70 GSH-Lac-modified peptides, 22 of which were also previously identified

as MeLac-alkyne probe-modified peptides (**Figure 3.11**; **Table iv and v, Appendix**). The search results also revealed a modification site preference shift. As shown in **Figure 3.12**, compared to MeLac-alkyne that modified proteins mostly on cysteine (more than 50% of modifications sites were cysteine), GSH-Lac modified proteins mostly on non-cysteine residues (less than 20% of modifications sites were cysteine). This discovery confirmed the proposed reaction routes of MeLac with amino acids, as MeLac and GSH-Lac differed significantly in the reaction route preference (the Michael addition reaction path for thiol with MeLac vs. acyl addition reaction path for amino and hydroxyl groups with GSH-Lac).

Modified peptides observed for glutathione S-transferase P1 (UniProt ID P09211, GSTP1) provided evidence for processes of in-cell assembly of GSH-Lac as well as its reactions with proteins (**Figure 3.10**; **Scheme 3.10**). Peptide YISLIYTNYEAGK together with its miscleaved counterpart YISLIYTNYEAGKDDYVK was observed with both MeLac-alkyne and GSH-Lac modifications at Y108; a natural variant[245] with 104I → V was also identified to carry the same modifications. Because the GSH-Lac modification initially appeared to be a mass shift of 843 Da on the 104I → V variant peptide from the unrestricted database search using only the canonical proteome sequence database (**Figure 3.10A**). The peptide ASCLYGQLPK, located in a narrow cavity (green, **Figure 3.10B**), was modified at C47 only by MeLac-alkyne. The N-terminal tryptic peptide PPYTVVYFPVR was only modified at Y7 by GSH-Lac. Y7 (red, **Figure 3.10B**) and Y108 (blue, **Figure 3.10B**) are located in a wide-open cavity. During the GSTP1 detoxification activity of xenobiotics, Y108 and Y7 are involved as electrophilic participants in the addition of GSH to exogenous molecules.[246] This catalytic mechanism is related to multidrug resistance in cancer therapy.[247] Anticancer prodrugs, which are activated by Y7, have thus targeted GSTP1.[248,249]

Therefore, a selective probe targeting GSTP1 has a direct impact in developing drugs regulating the protein activity.

The formation of observed GSTP1 adducts with MeLac-alkyne and GSH-Lac could be explained by different routes, as shown on **Scheme 3.11**. The first route initiated from MeLac-alkyne reacting with GSH-bound GSTP1 to form the MeLac-alkyne GSTP adduct at Y108 (and C47) through steps A and B; additional probe modifications on nearby residues were also observed (**Figure 3.12**). Further attachment of GSH to the Y108 adduct formed the same product as the directed reaction product with GSH-Lac (step D and E, **Scheme 3.11**). Consequently, the Y108 peptides were detected with both MeLac-alkyne and GSH-Lac modifications. The formation of GSH-Lac or in-live-cell installation of GSH to MeLac-alkyne was likely to be catalyzed by GSTP1. When GSH-Lac reacted locally with GSTP1, both Y108 and Y7 were modified (steps E and F, **Scheme 3.11**). It was also plausible that newly formed GSH-Lac could diffuse away from GSTP1 and react with other intracellular proteins (step G and H, **Scheme 3.11**), resulting in the detection of 48 modified peptides unique to GSH-Lac. In contrast to the reactions between MeLac-alkyne and its protein targets corresponding to the identified 640 unique peptides, the reaction between GSH-Lac and its protein targets had substantial selectivity, which was indicated by it smaller target pool of proteins corresponding to only 48 unique modified peptides. (**Figure 3.11**; also see **Table v on Appendix** for more details of these peptides)

GSH-Lac, not MeLac-alkyne modified Y7. The accessibility of Y7 and Y108 to the β-lactone C2 (acyl) carbon on GSH-Lac was evaluated via computational protein-ligand docking using Schrödinger Suite. Up to five energetically favorable protein-ligand interacting models were calculated for all four possible β-lactone stereoisomers of GSH-Lac. The GSH-Lac docked protein models showed stereoisomer (*3S,4S*) and (*3S,4R*) stereoisomers of GSH-Lac had relatively lower

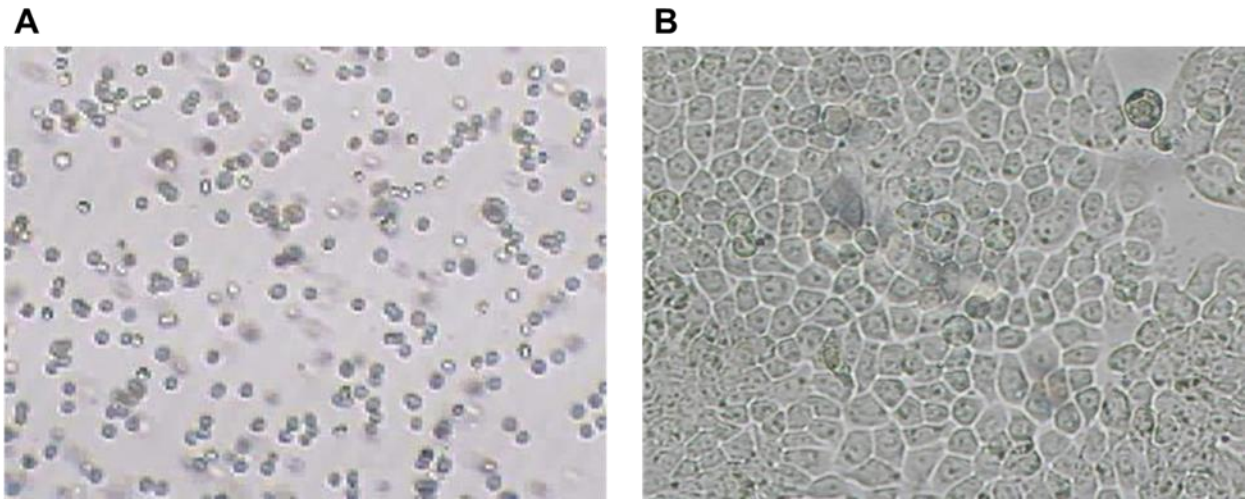
docking energy (Glide energy) and closer distances to the oxygen atoms of both Y7 and Y108 compared to the other two stereoisomers. These docking results implied possible stereoselectivity on GSH-Lac by the GSTP1 catalytic cavity featuring Y7 and Y108 (**Figure 3.14**). Comparing (*3S,4S*) and (*3S,4R*), the latter was docked with the shortest distances to both Y7 and Y108 (red projection dots highlighted in the green ellipse, **Figure 3.14A**), while the former had slightly longer distances to Y108 (blue projection dots highlighted in the yellow ellipse, **Figure 3.14A**) but the lowest Glide energy. This observation suggested Y7 favored the reaction with (*3S,4S*), but Y108 was closed to (*3S,4R*). Nevertheless, validation of the GSTP1 stereoselectivity on GSH-Lac would require future dedicated experiments using stereo-purified synthetic GSH-Lac.

3.4 Conclusion

MeLac is a novel chemical probe warhead that couples a Michael acceptor with a β-lactone moiety. This small, rigid warhead's broad reactivity allows it to react with different protein nucleophiles through distinct mechanisms. Multiple reactive sites on MeLac extend the scope of these reactions and provide a potential means of conjugating recognizing moieties for ligand-directed chemistry[250] in live cells. With its broad reactivity, the MeLac warhead also offers a unique possibility to convert non-covalent inhibitors to measurement probes targeting different enzymes,[251] as well as to broaden druggable targets by identifying new binding sites.[77,78] MeLac probes, being useful for live-cell applications, potentially have an immense impact on drug discovery and further expand the analysis of the reactions with proteins that can deactivate during cell lysis. MeLac-based competitive ABPP platforms are highly adaptable and are appropriate for measuring the live-cell action of a wide array of small reactive molecules.

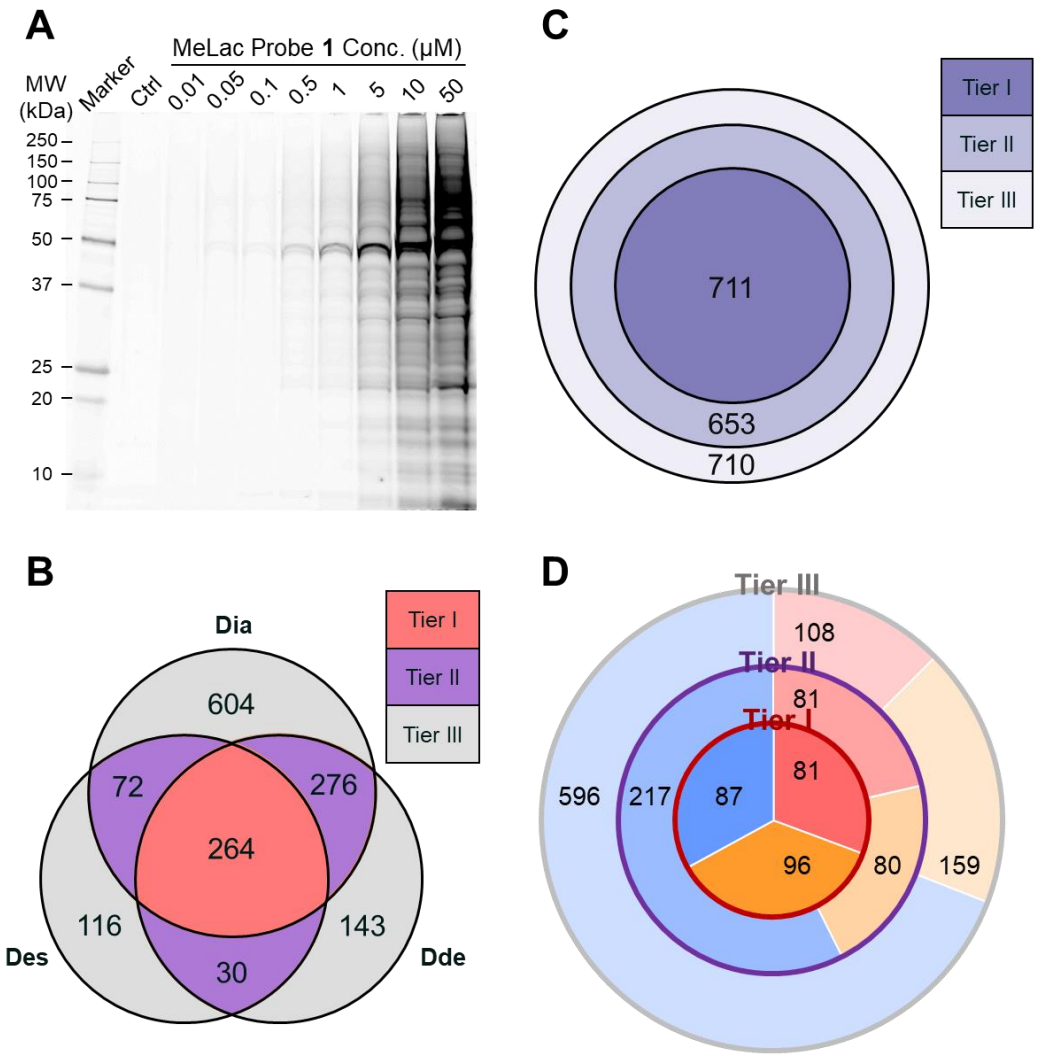
3.5 Chapter 3 Figures

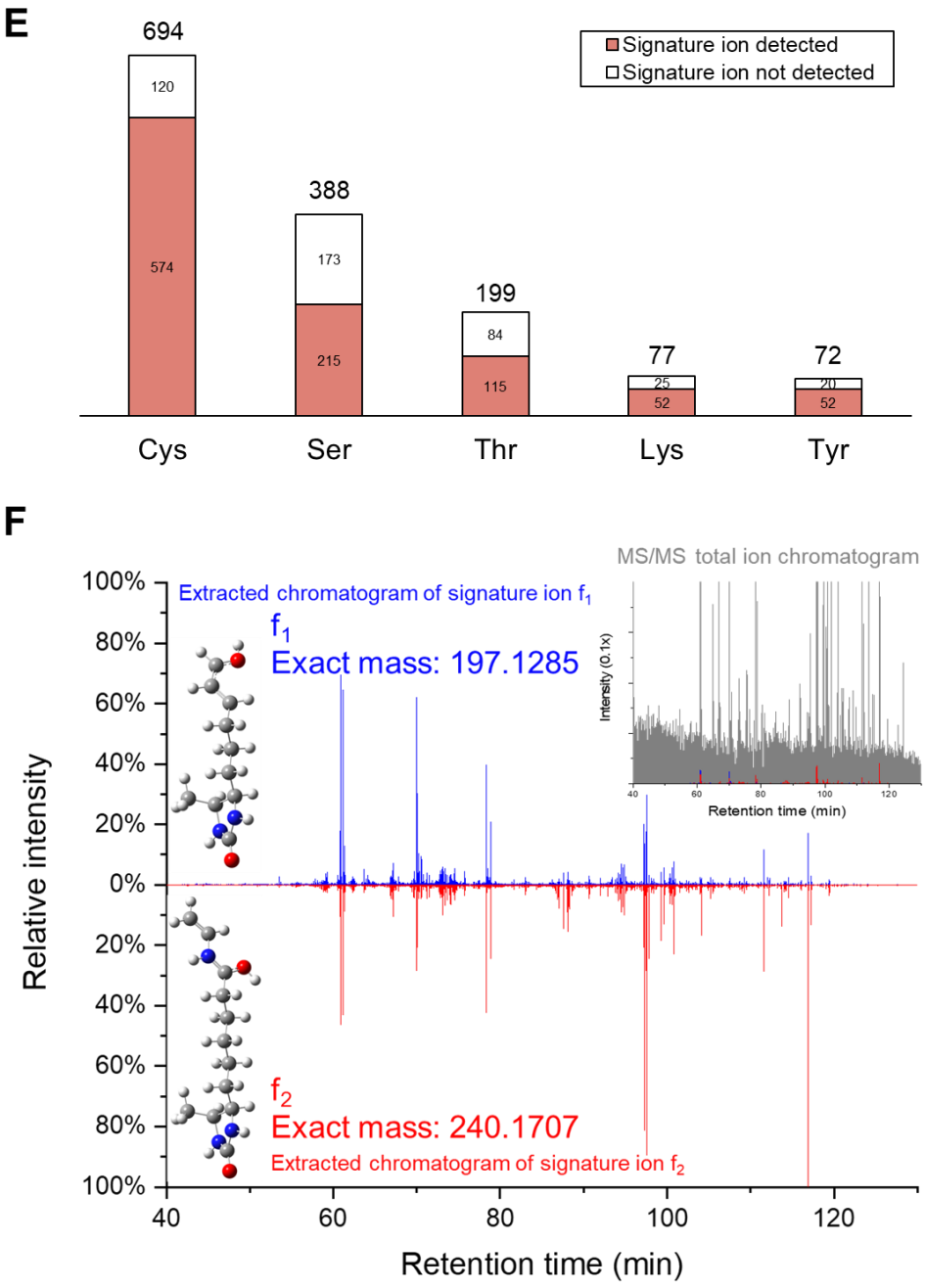
Figure 3.1 HT-29 cells under microscope.



Note: (A) Seeds and **(B)** Target confluency (90 to 95%).

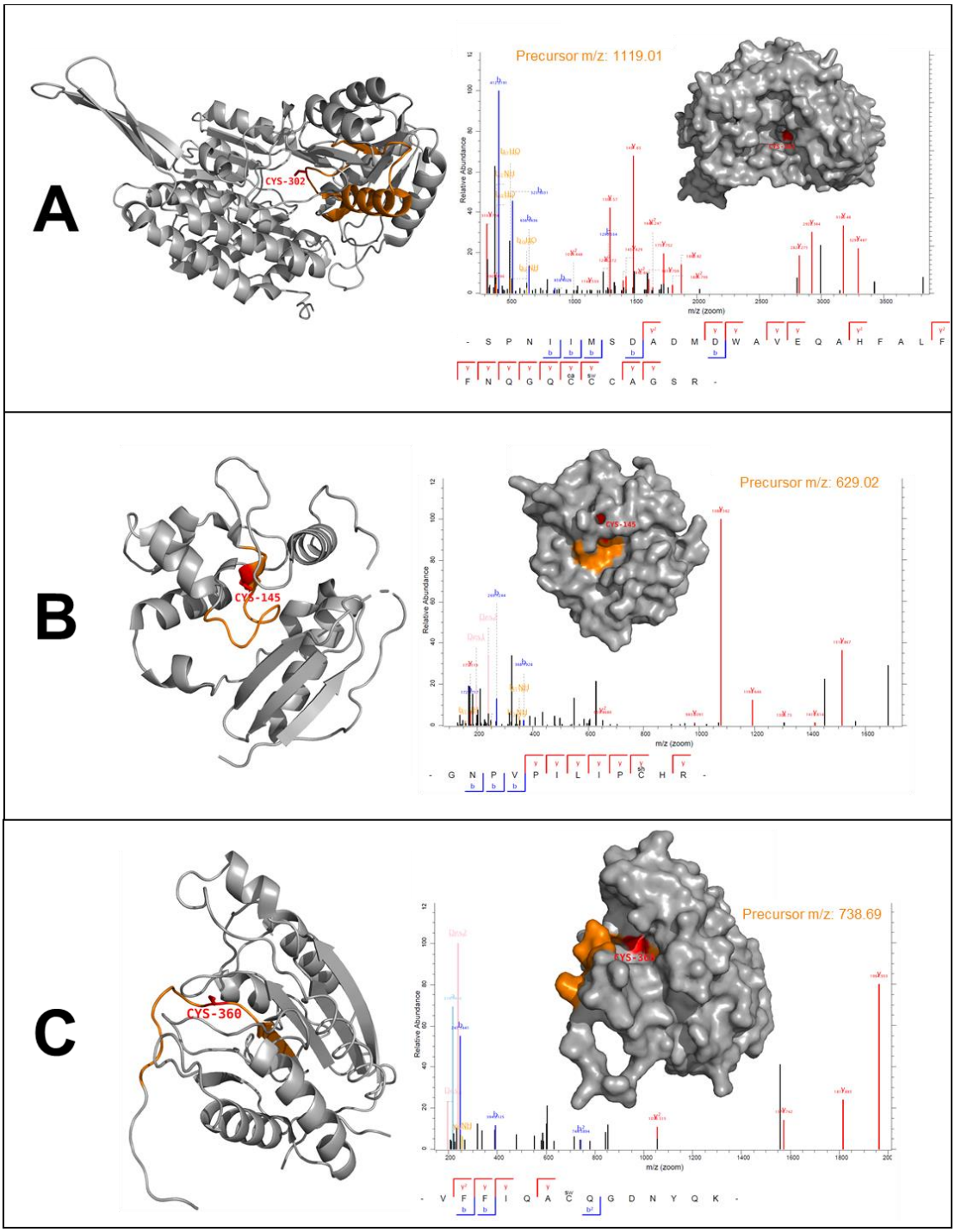
Figure 3.2 Multi-level characterization of protein reactions with α-methylene-β-lactone (MeLac) alkyne probe.





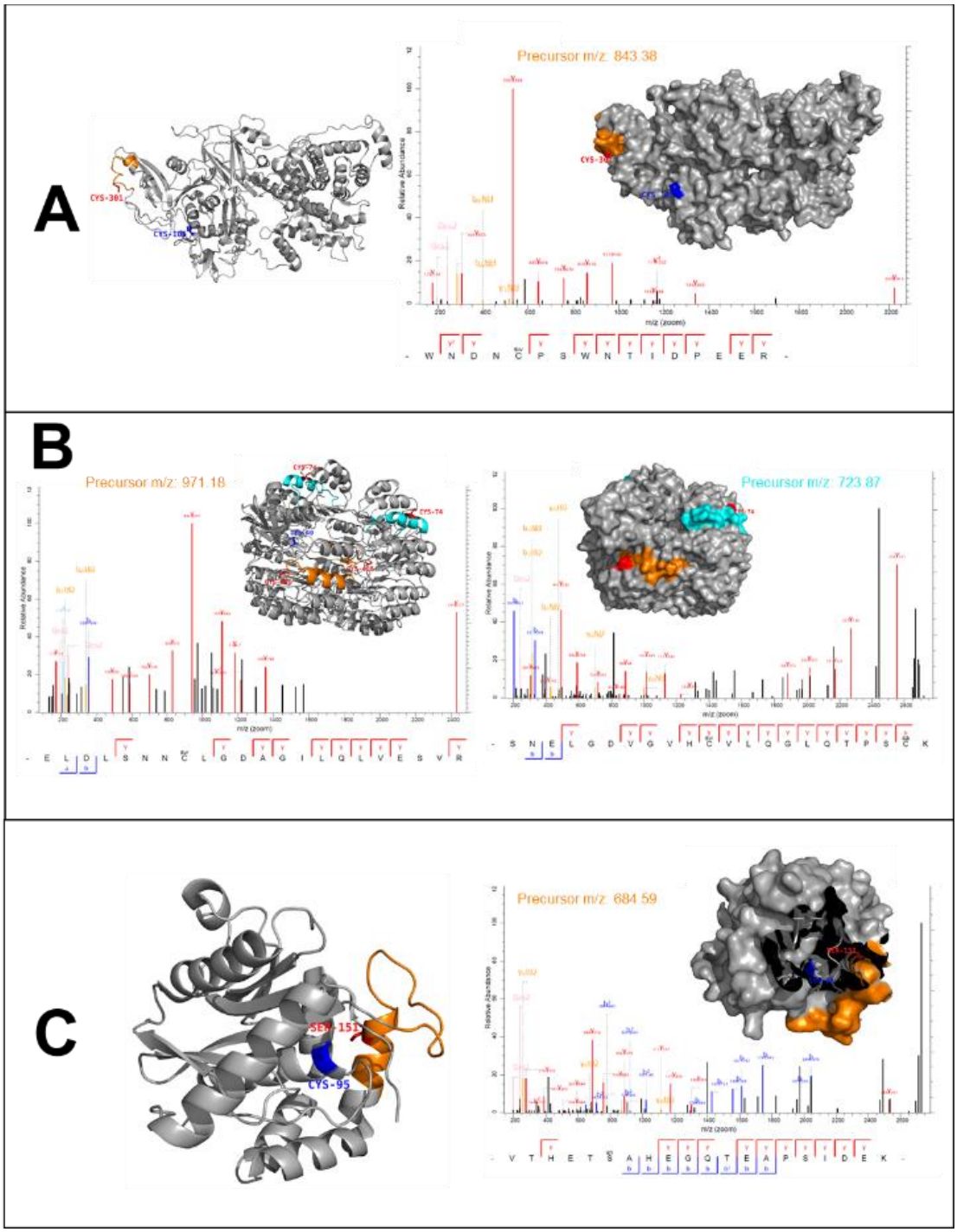
Note: (A) Gel-based activity-based protein profiling showing broad reactivity of the MeLac-alkyne probe. **(B)** Triplicated selective enrichment (Affinity Tagging Triplication) of probe-reacted proteins resulting in discrete identification confidence (Tier I > Tier II > Tier III) of probe-reacted proteins annotated with gene names. **(C)** Three-tier identification confidence ranking of isoform-specific proteins annotated with UniProt accession numbers. **(D)** Comparison of identification numbers of protein-level vs. peptide-level enrichment. **(E)** MeLac modification sites on nucleophilic amino acid residues identified at a residue-level resolution from peptide-level enrichment workflow. **(F)** Signature ion-based reduction of data complexity and validation of identification of probe-modified peptide; insert comparing total ion chromatogram of peptide fragment ions and extracted ion chromatograms for highly specific fragment ions of desthiobiotin-PEG_3 (f1 and f2; structures of minimized geometry with the lowest total energy shown).

Figure 3.3 Example protein models showing probe-modified catalytical residues.



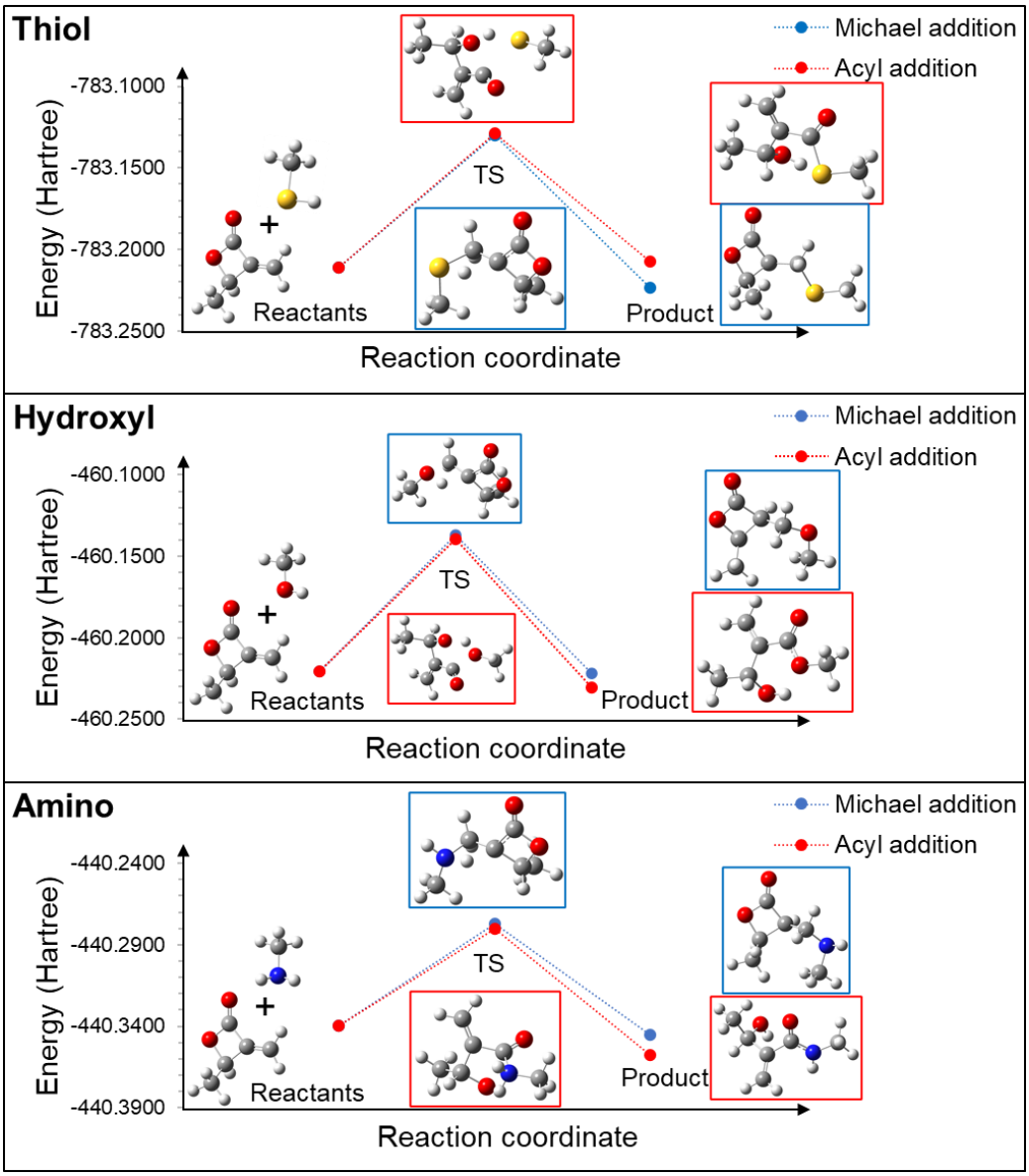
Note: The identified/sequenced probe-modified peptides are highlighted in orange. The probe-modified residues are highlighted in red. **(A)** Aldehyde dehydrogenase, mitochondrial (ALDH2, PDB ID: 1o04).[252] **(B)** Methylated-DNA-protein-cysteine methyltransferase (MGMT, PDB ID: 1eh6).[253] **(C)** Caspase-8 (CASP8, PDB ID: 1qtn).[254]

Figure 3.4 Example protein models showing probe-modified non-catalytic residues.



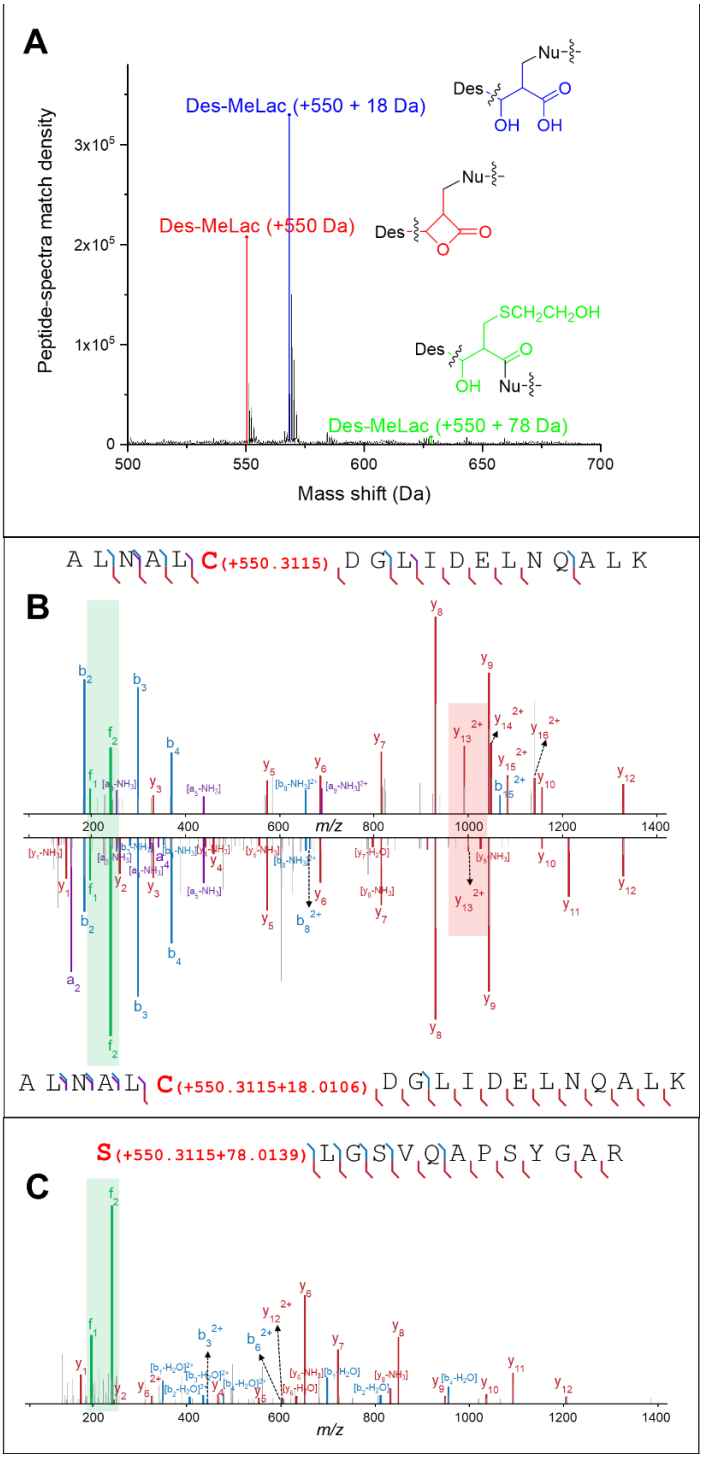
Note: The identified/sequenced probe-modified peptides are highlighted in orange (or cyan). The probe-modified residues are highlighted in red. The reported catalytical residues are highlighted in blue. **(A)** Calpain-2 catalytic subunit (CAN2, PDB ID: 1kfu).[255] **(B)** Ribonuclease inhibitor (RINI, PDB ID: 1a4y).[256] **(C)** Ubiquitin carboxyl-terminal hydrolase isozyme L3 (UCHL3, PDB ID: 6qml).

Figure 3.5 Theoretical MeLac reaction paths with nucleophiles.



Michael addition (kcal/mol)		
	$\Delta G_{(TS-Reactants)}$	$\Delta G_{(Product-Reactants)}$
Thiol	50.9	-7.9
Hydroxyl	52.6	-0.6
Amino	39.4	-3.7
Acyl addition (kcal/mol)		
	$\Delta G_{(TS-Reactants)}$	$\Delta G_{(Product-Reactants)}$
Thiol	51.8	2.4
Hydroxyl	51.0	-6.1
Amino	37.3	-11.4

Figure 3.6 Global and residue-specific illustration of MeLac modifications on peptides.



Note: (A) A DeltaMass plot showing identified global modification of peptides with mass shifts accountable to the conjugation with MeLac. **(B)** Representative MeLac modifications on peptides at Cys residue. **(C)** Representative MeLac modification on peptides at Ser residue.

Figure 3.7 Venn diagram comparison of MeLac vs. IA alkyne[77] probe-modified tryptic peptides.

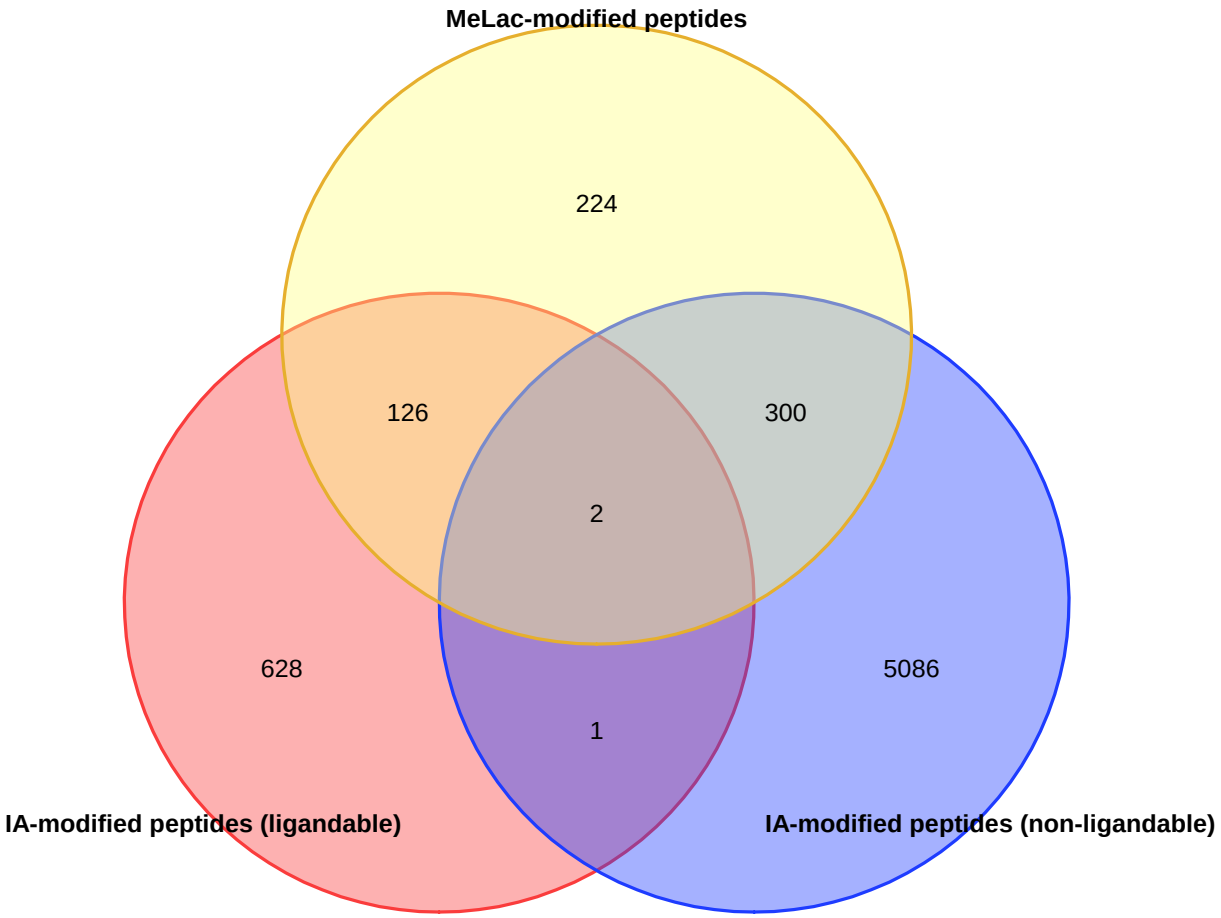
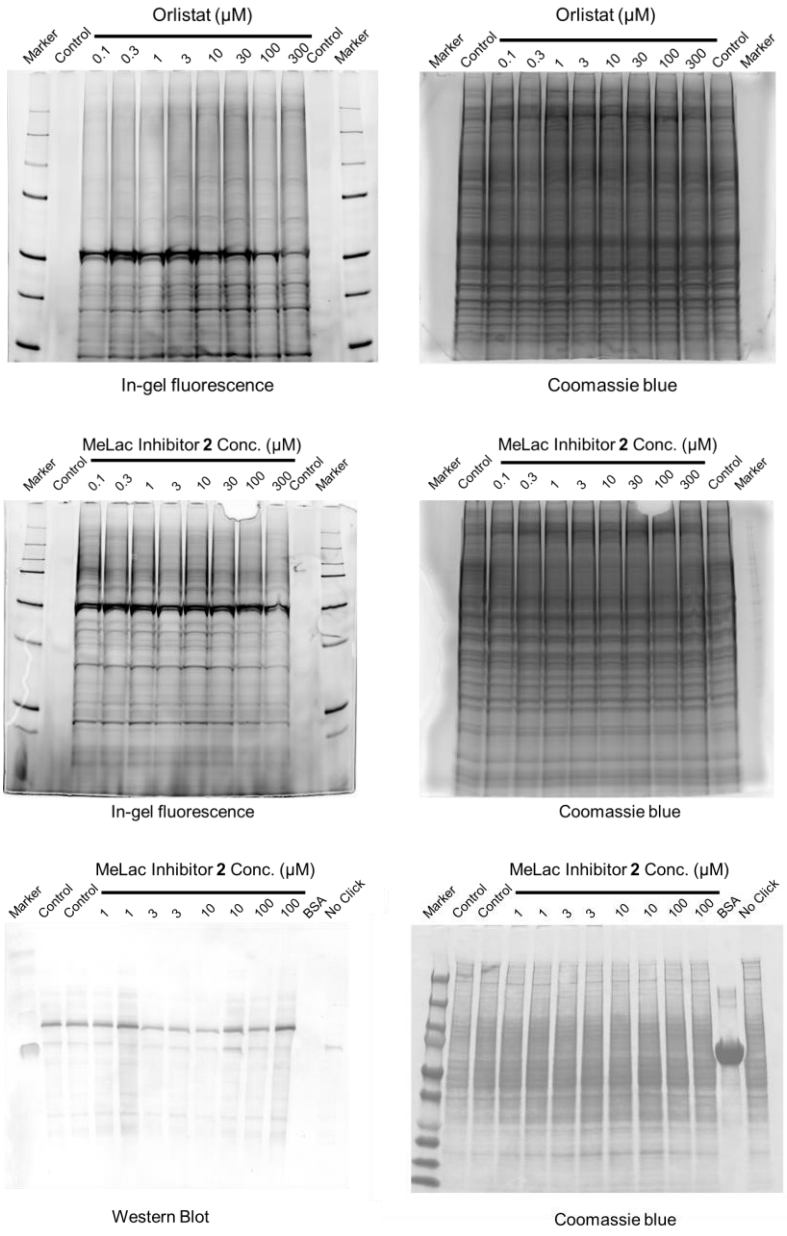
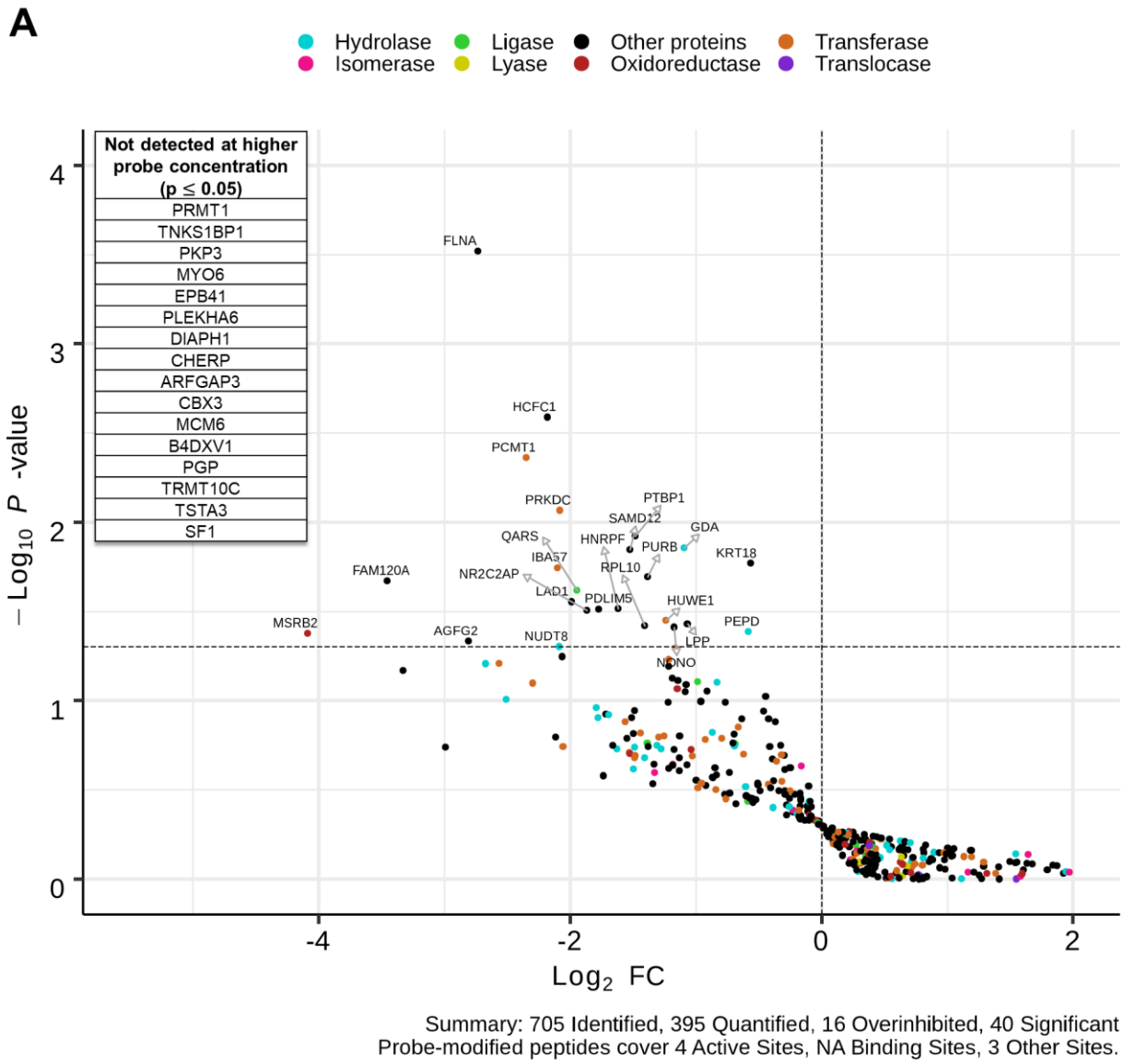


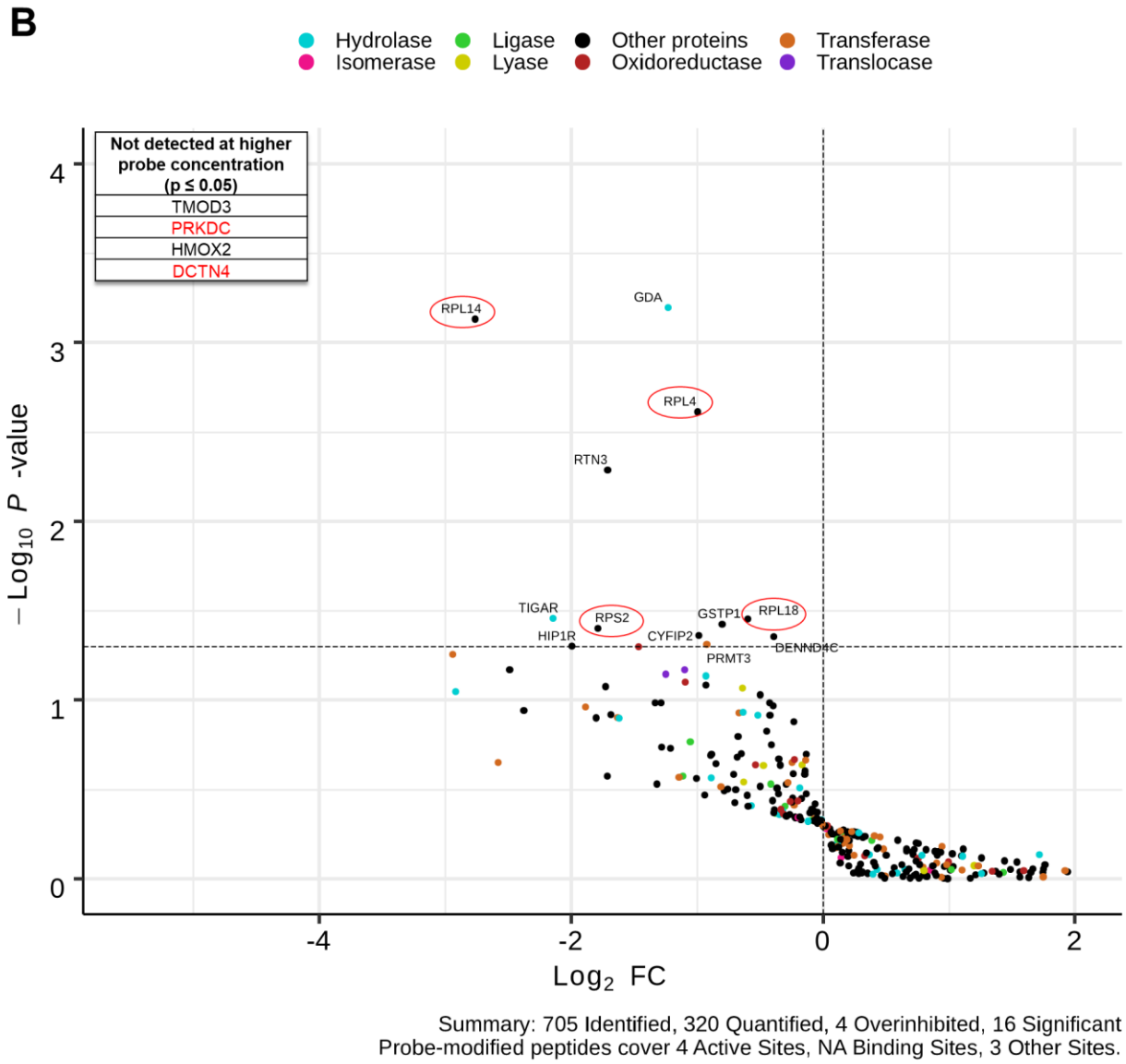
Figure 3.8 Competitive activity-based protein profiling displaying broad reactivity of the alkyl MeLac inhibitor.



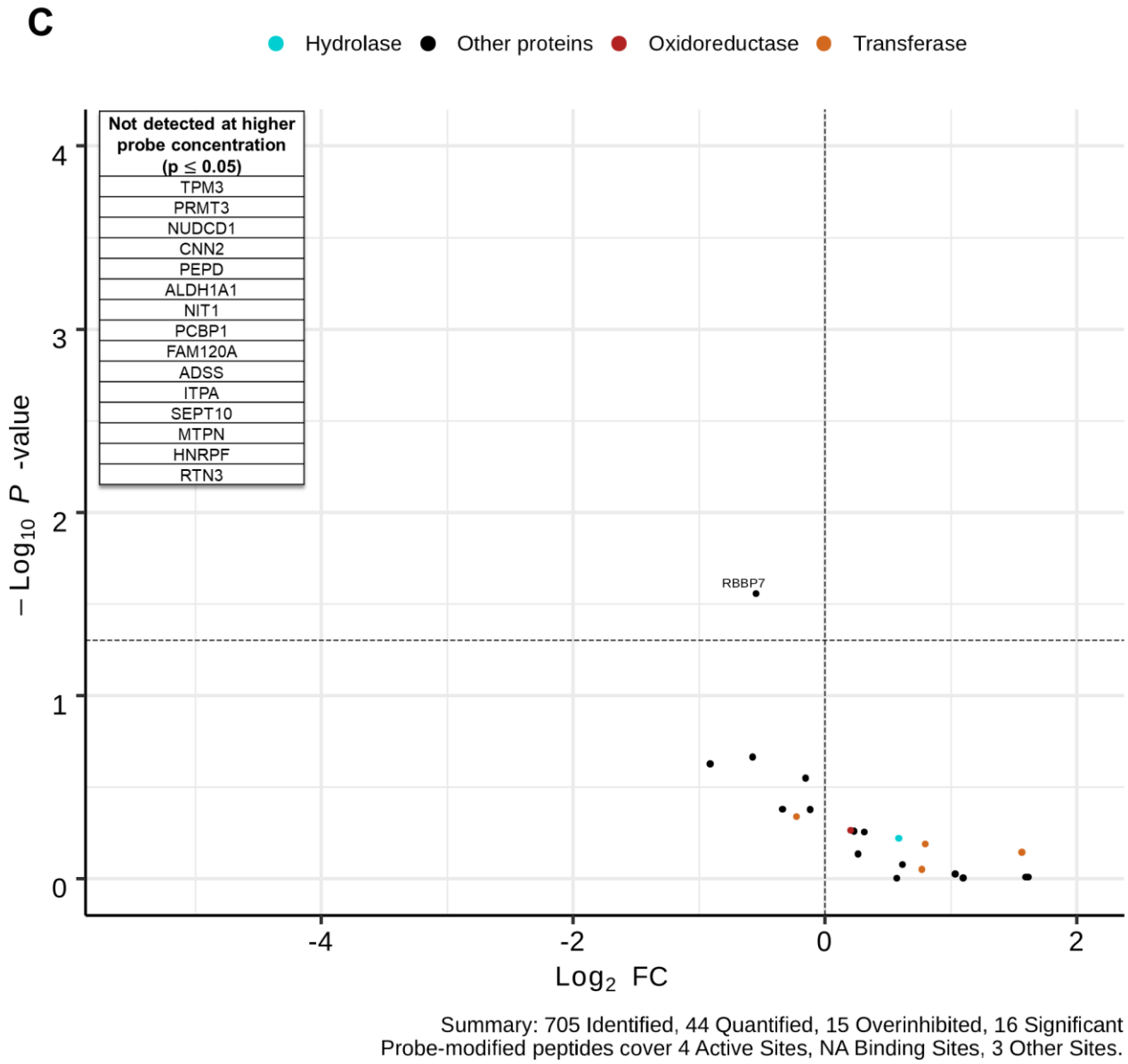
Note: HT-29 Cells were first incubated with the alkyl MeLac inhibitor or orlistat at the corresponding concentration for 30 minutes. Cells were then incubated for 30 minutes with the MeLac-alkyne probe at 5 μM for 1 hour. Cells were washed and lysed. The subsequent lysate underwent CuAAC with TAMRA biotin azide as the fluorescent reporting tag or Des as the affinity tag. (See **Experimental 3.2.5** for detailed procedure) Approximately 100 μg of total protein was loaded to each well without affinity enrichment. Gel images were acquired to show fluorescent protein bands (probe-reacted proteins) and Coomassie Blue-stained protein bands (all proteins).

Figure 3.9 MeLac-alkyne quantitively probing protein selectivity/reactivity profiles of chemically distinct molecules.



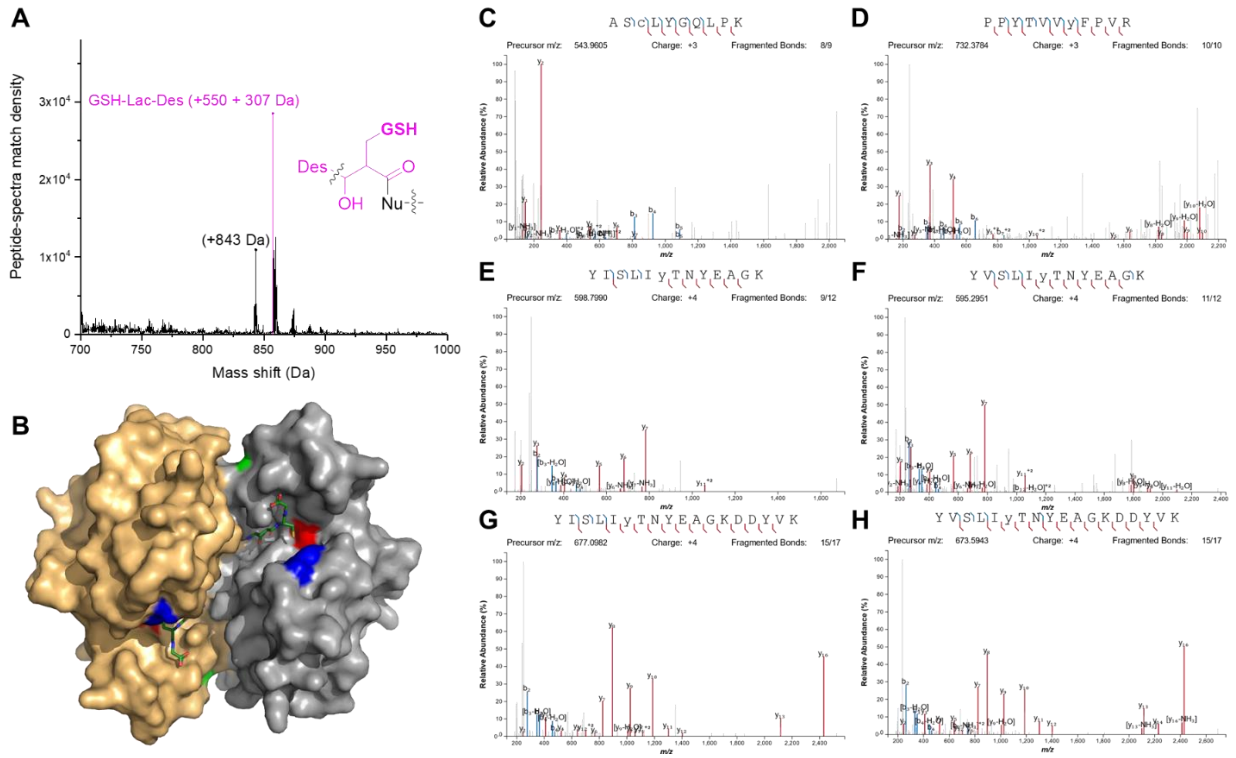


Summary: 705 Identified, 320 Quantified, 4 Overinhibited, 16 Significant
Probe-modified peptides cover 4 Active Sites, NA Binding Sites, 3 Other Sites.



Note: The volcano plots show that (**A**) orlistat, (**B**) parthenolide, and (**C**) alkyl MeLac selectively reacted with different proteins. These reactions were detectable by changing the inhibitor concentrations, 10 vs. 1 μM for orlistat and parthenolide, 100 vs. 10 μM for alkyl MeLac, using the MeLac-alkyne as the measurement probe. Inert tables showing proteins of quantified probe-modified peptides at the lower inhibitor concentration with p-value ≤ 0.05, but not detected at the higher inhibitor concentration (zero intensity for all samples at the higher inhibitor concentration).

Figure 3.10 Identification of GSH-Lac adducts.



Note: (A) The DeltaMass plot showing identified global modification of peptides with mass shifts accountable to the conjugation with GSH-Lac. **(B)** GSH-bound GSTP1 dimmer (PDB ID: 5j41)[231] with three reactive residues labeled: Y7 in red, Y108 in blue, and C47 in green. MS/MS spectra for a GSTP1 peptide with modifications of MeLac-alkyne uniquely at C47 **(C)**, GSH-Lac uniquely at Y7**(D)**, and peptides with modifications of both MeLac-alkyne and GSH-Lac detected at Y108 **(E** and **F)**. Spectra for peptides of a mutant GSTP1 isoform carrying modifications of both MeLac-alkyne and GSH-Lac detected at Y108 **(G** and **H)**.

Figure 3.11 Selectivity comparison of GSH-Lac vs. MeLac-alkyne at the peptide level.

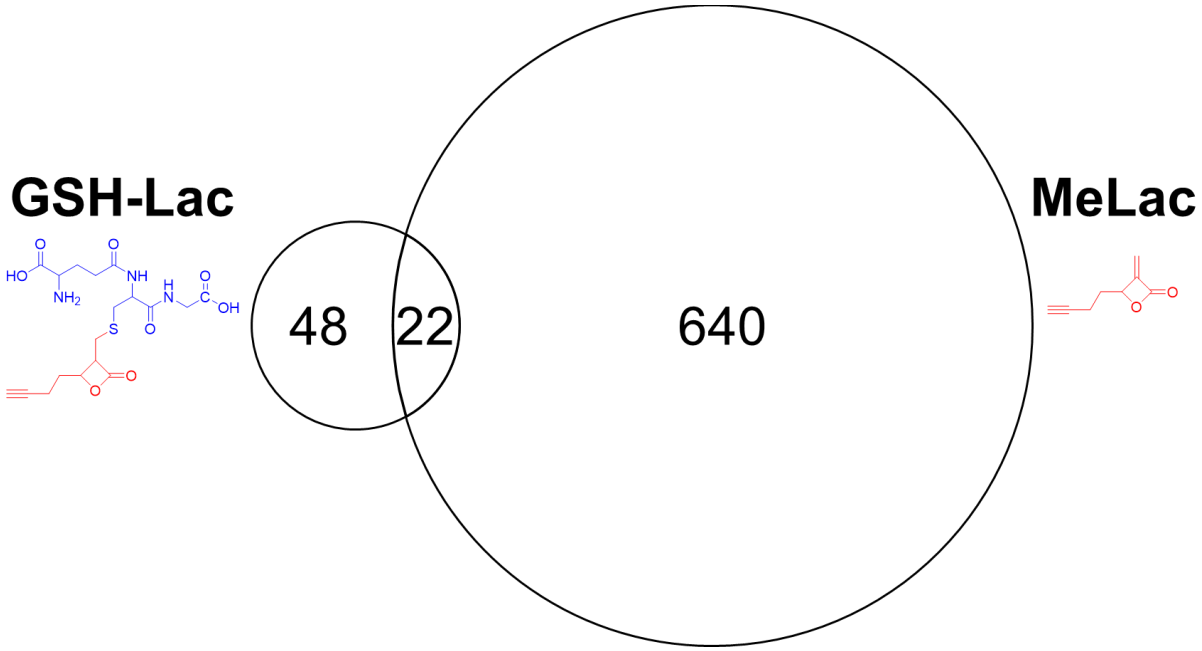


Figure 3.12 Summary of probe modification sites.

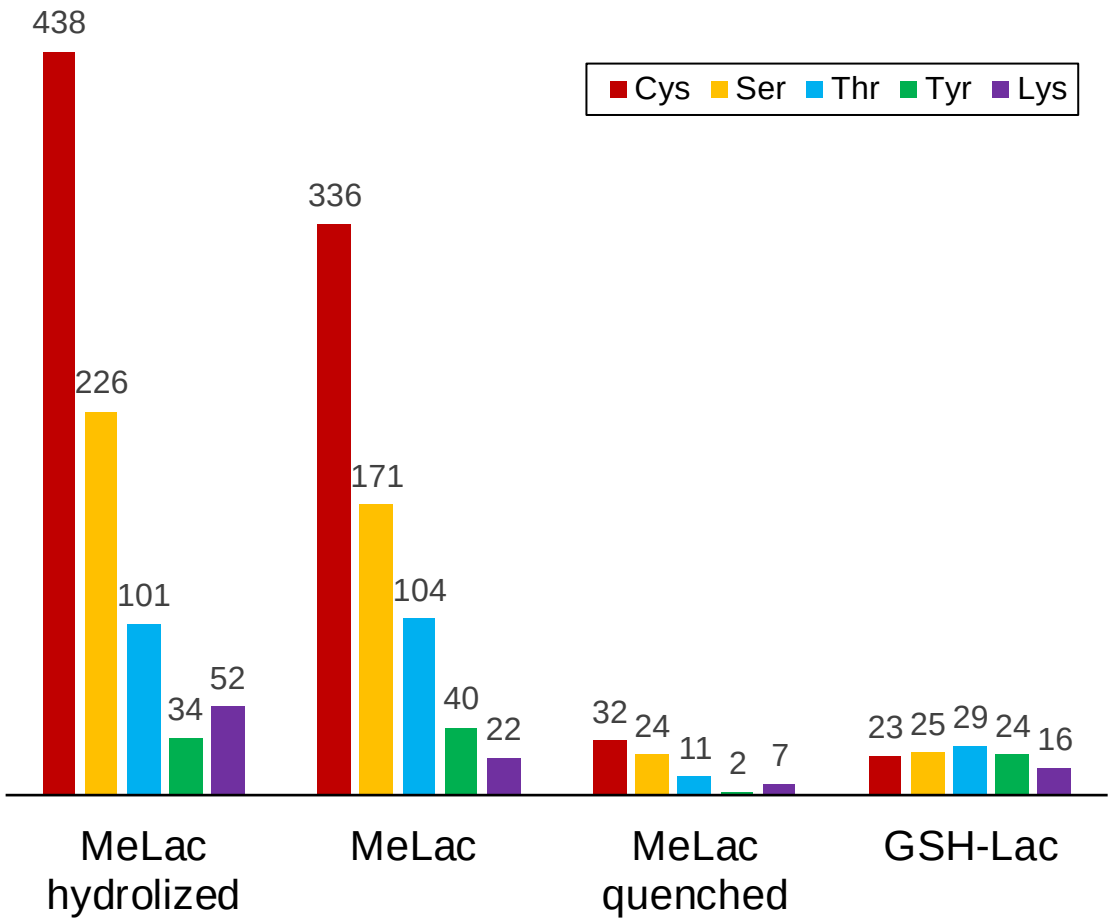
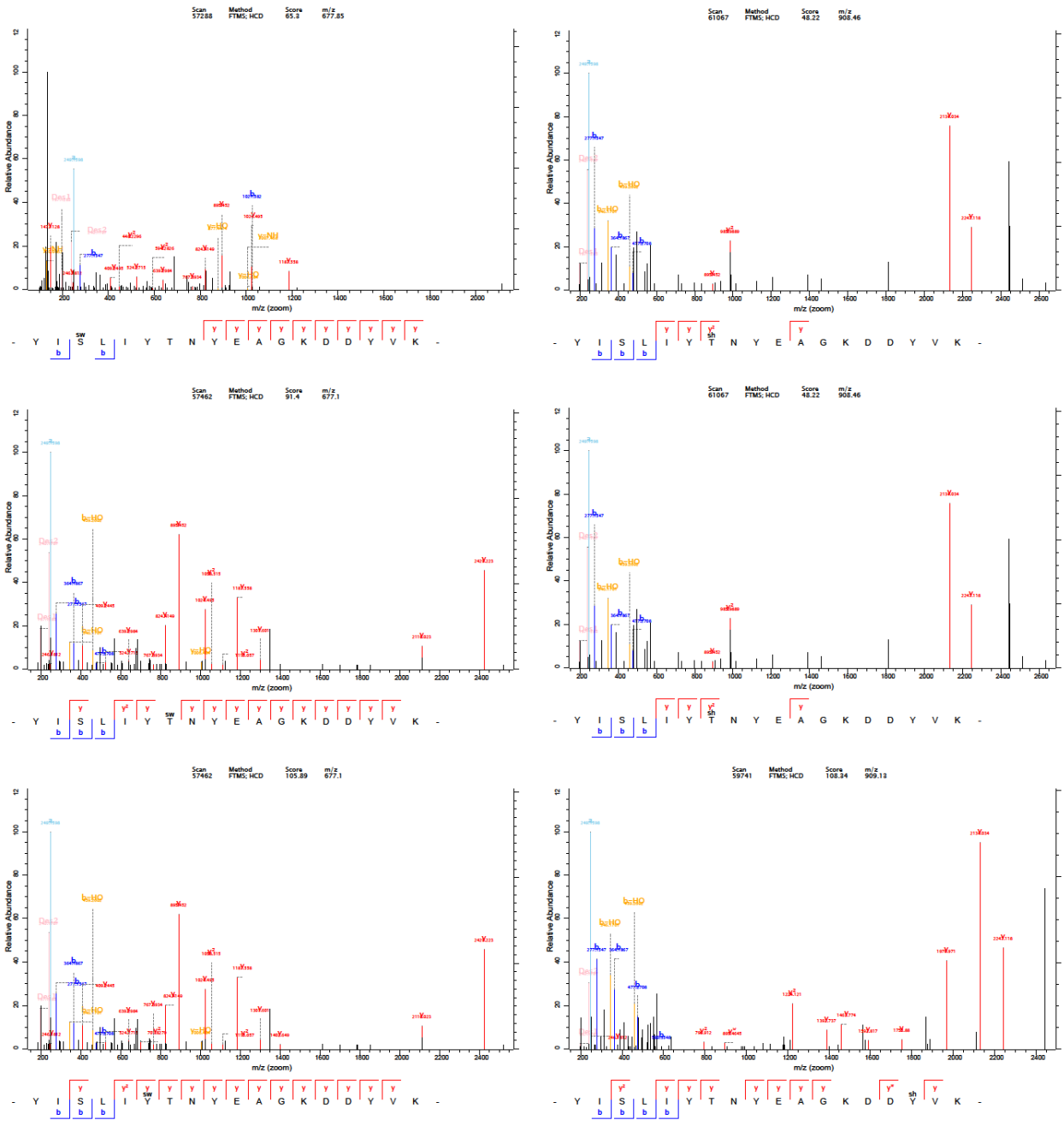
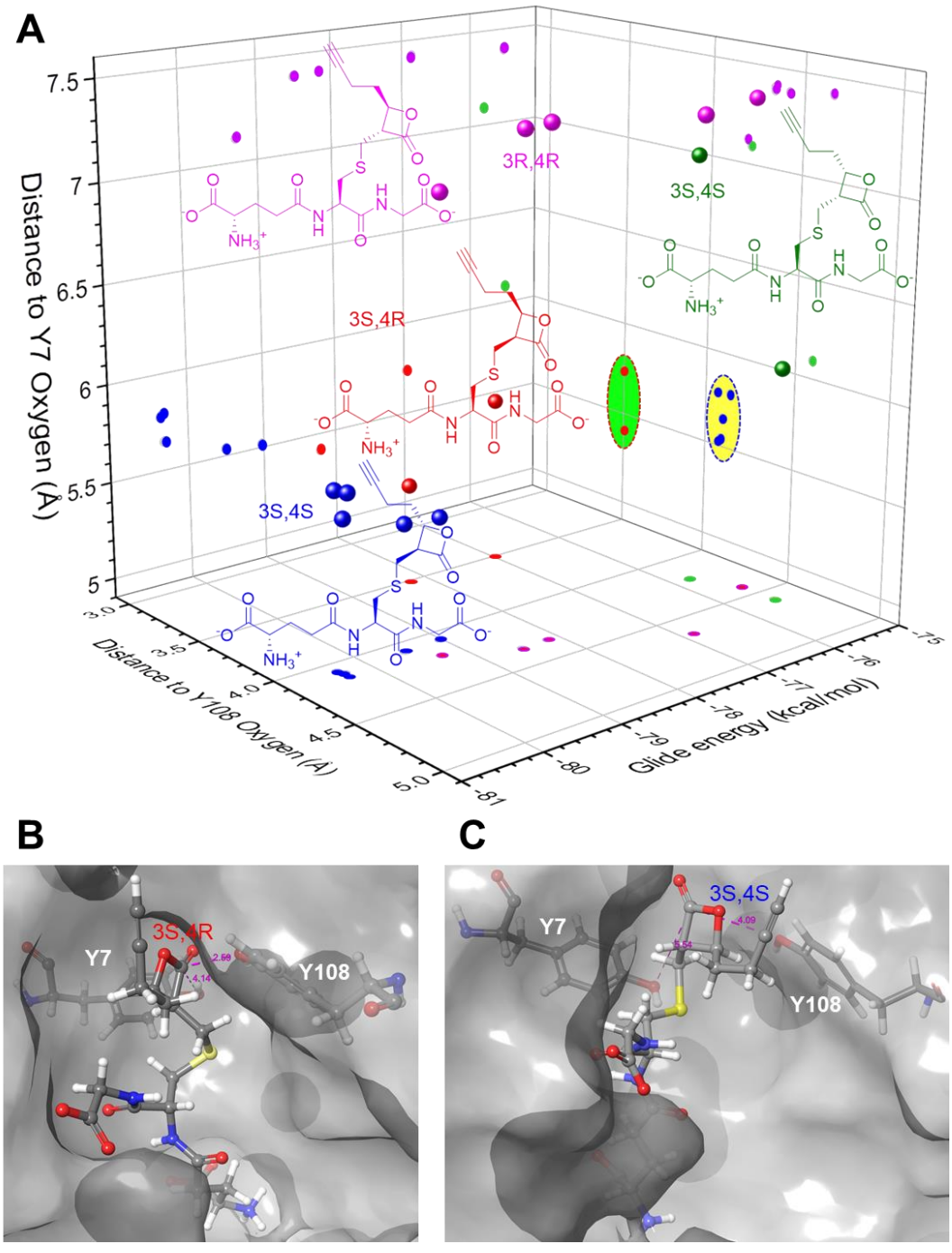


Figure 3.13 Additional MS/MS spectra observed different MeLac modification sites on peptide YISLIYTNYEAGKDDYVK of glutathione S-transferase P.



Note: sw = Des MeLac **w**arhead modification (+ 550.3115 Da), sh = Des MeLac **h**ydrolyzed modification (+ 568.3221 Da).

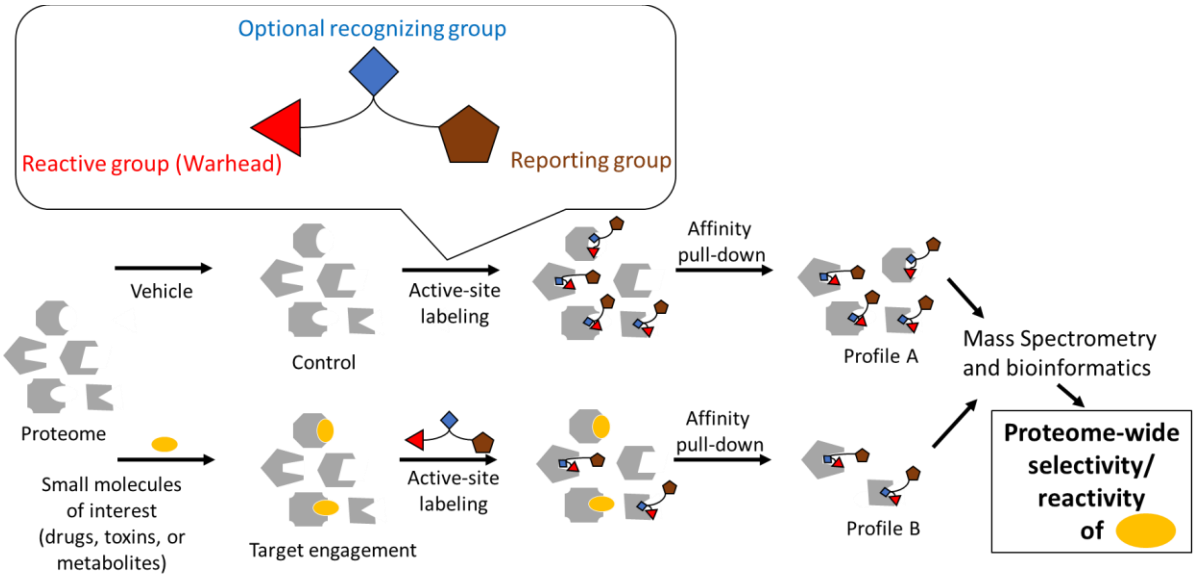
Figure 3.14 MeLac-alkyne recruiting endogenous glutathione to assemble a selective β-lactone probe.



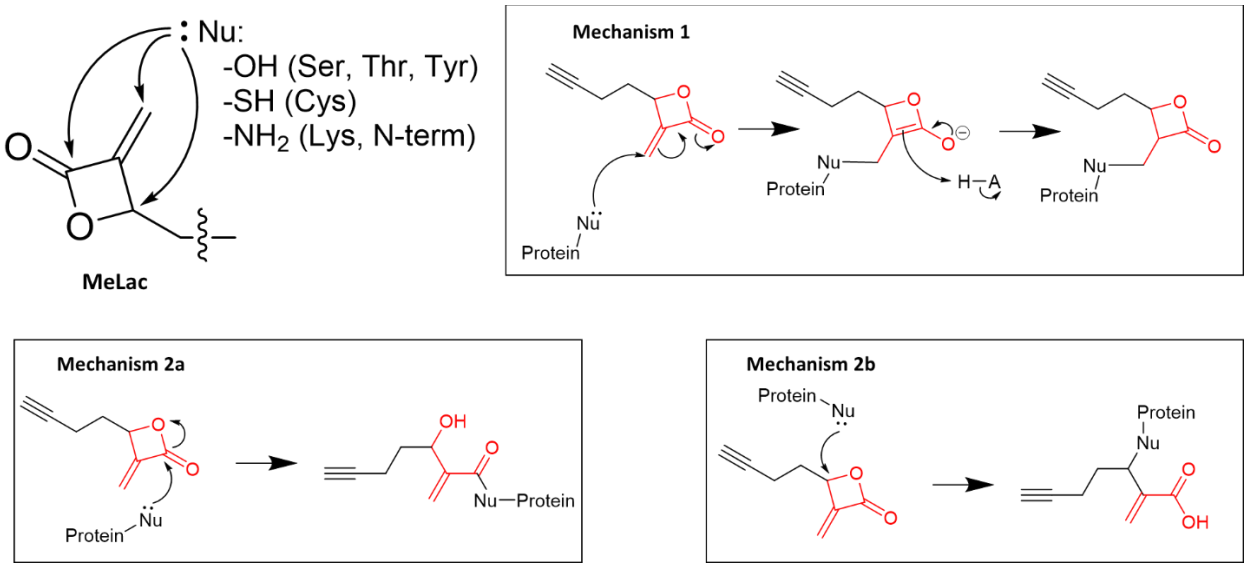
Note: (A) Selectivity comparison, ID numbers of probe-modified peptides, GSH-Lac vs. MeLac-alkyne **(B)** Minimized docking pose of GSH-Lac probe in GSTP1 for selective reaction with Y108 with possible general acid catalysis by Y7. **(C)** Docking energy comparison indicating differentiation of GSH-Lac stereoisomers by the reaction site of GSTP1.

3.6 Chapter 3 Schemes

Scheme 3.1 Illustration of the "3-R" anatomy of an activity-based probe and general design an activity-based protein profiling experiment.

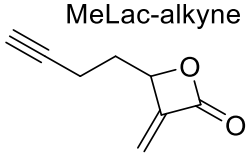


Scheme 3.2 Proposed reactivity of MeLac towards different protein nucleophiles via distinct mechanisms.



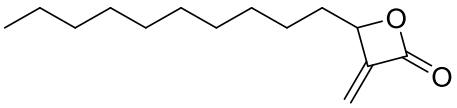
Scheme 3.3 Structures of MeLac-alkyne probe, alkyl MeLac inhibitor, orlistat, and parthenolide.

MeLac-alkyne probe



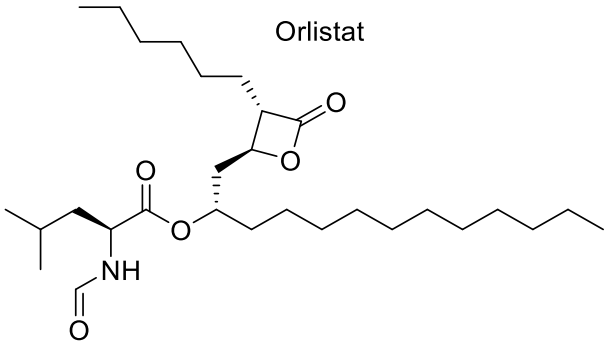
Chemical Formula: $C_8H_8O_2$
Exact Mass: 136.0524
Molecular Weight: 136.1500

Alkyl MeLac inhibitor



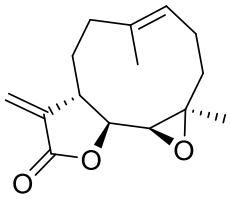
Chemical Formula: $C_{14}H_{24}O_2$
Exact Mass: 224.1776
Molecular Weight: 224.3440

Orlistat



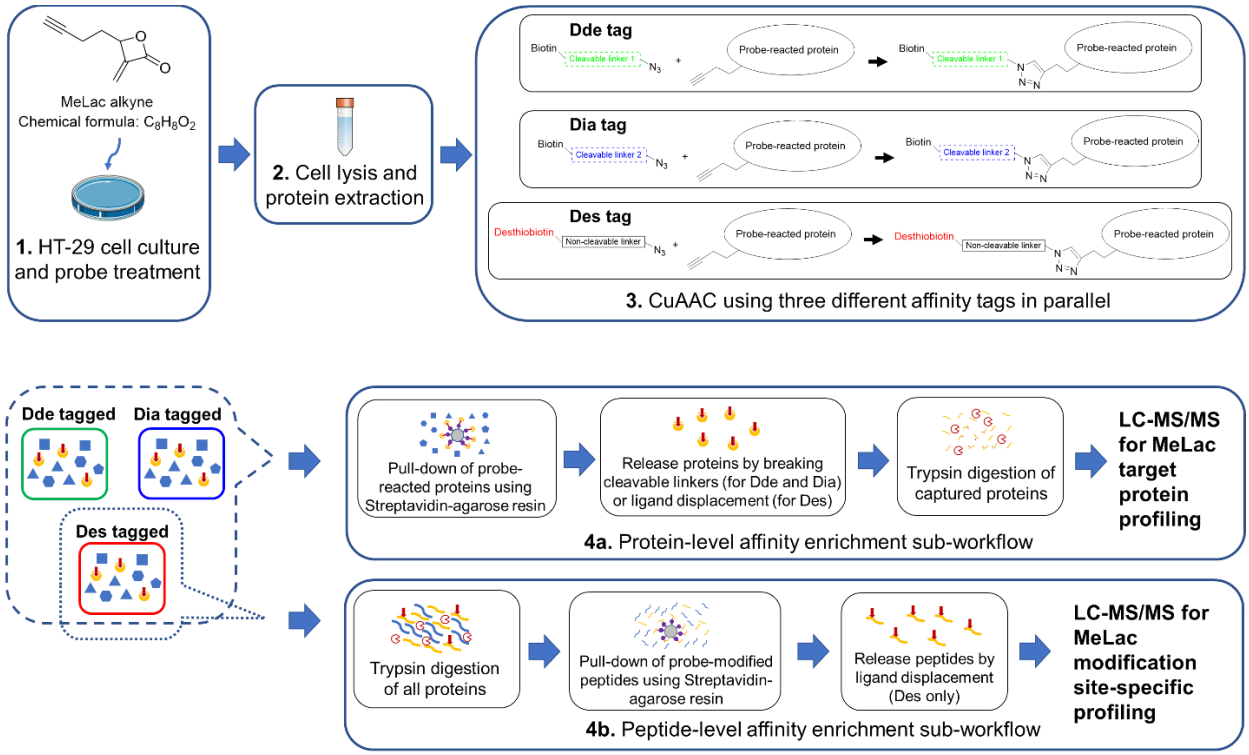
Chemical Formula: $C_{29}H_{53}NO_5$
Exact Mass: 495.3924
Molecular Weight: 495.7450

Parthenolide

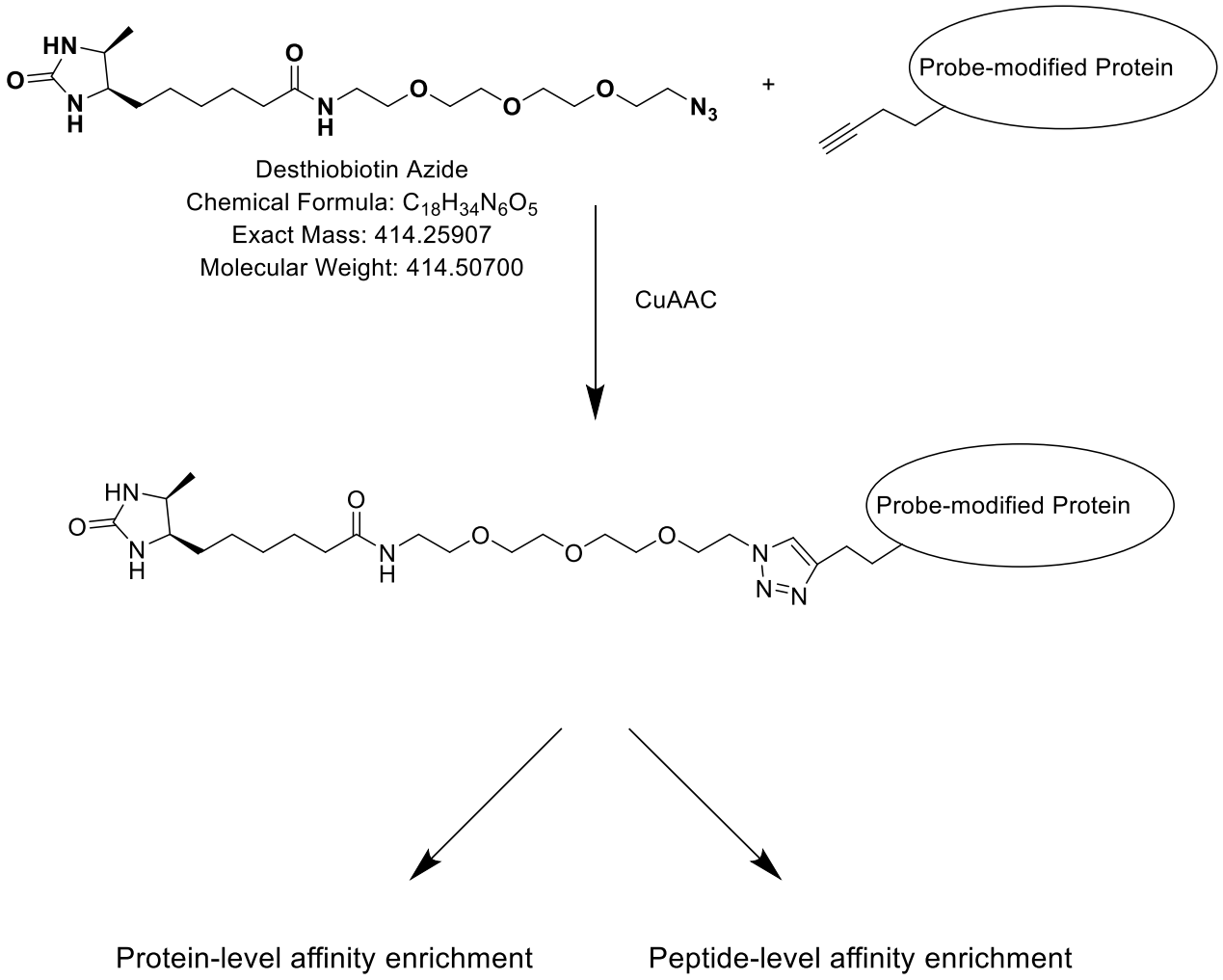


Chemical Formula: $C_{15}H_{20}O_3$
Exact Mass: 248.1412
Molecular Weight: 248.3220

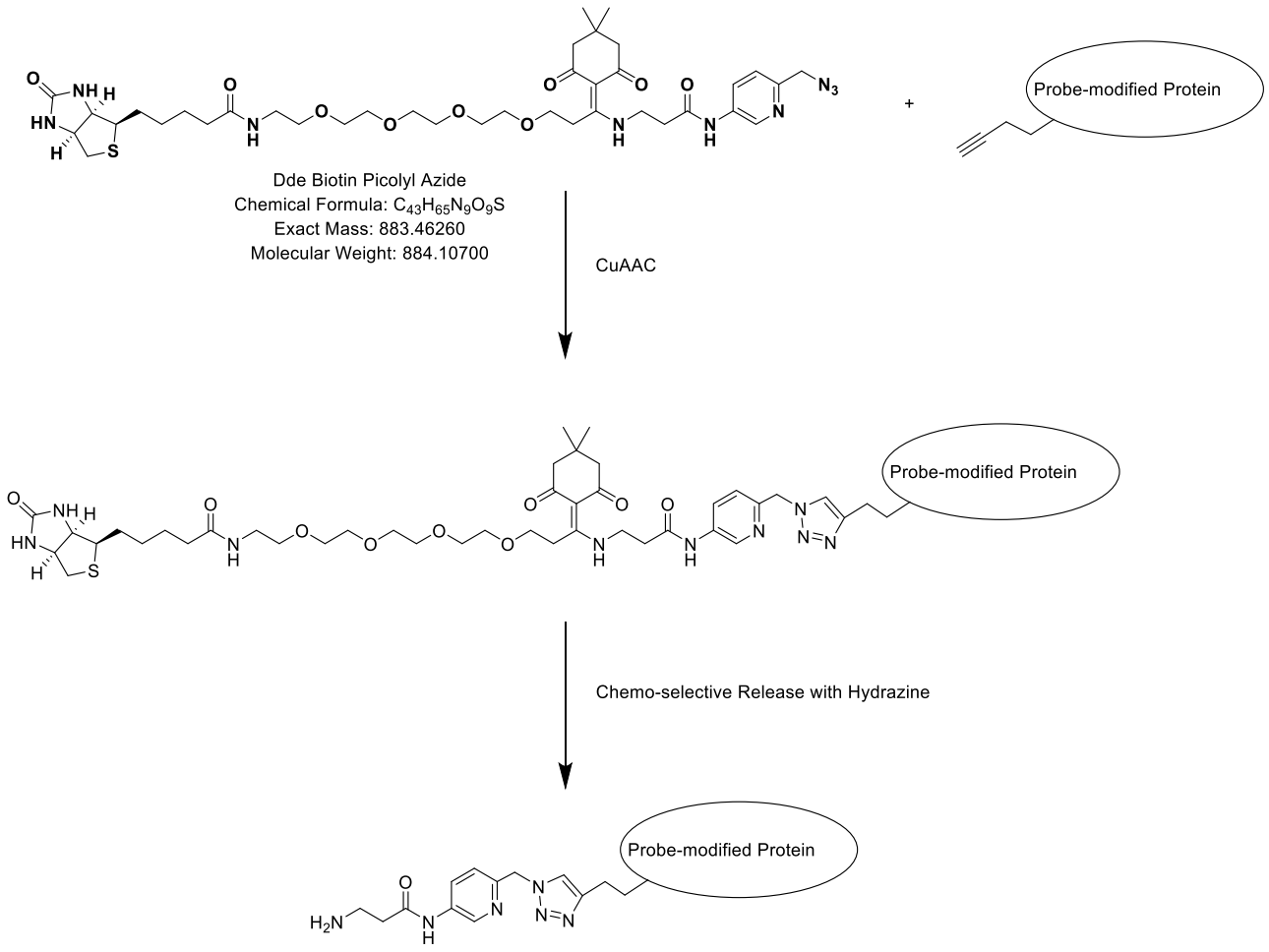
Scheme 3.4 Overview of the experimental workflow featuring affinity tagging triplication and dual-level enrichment



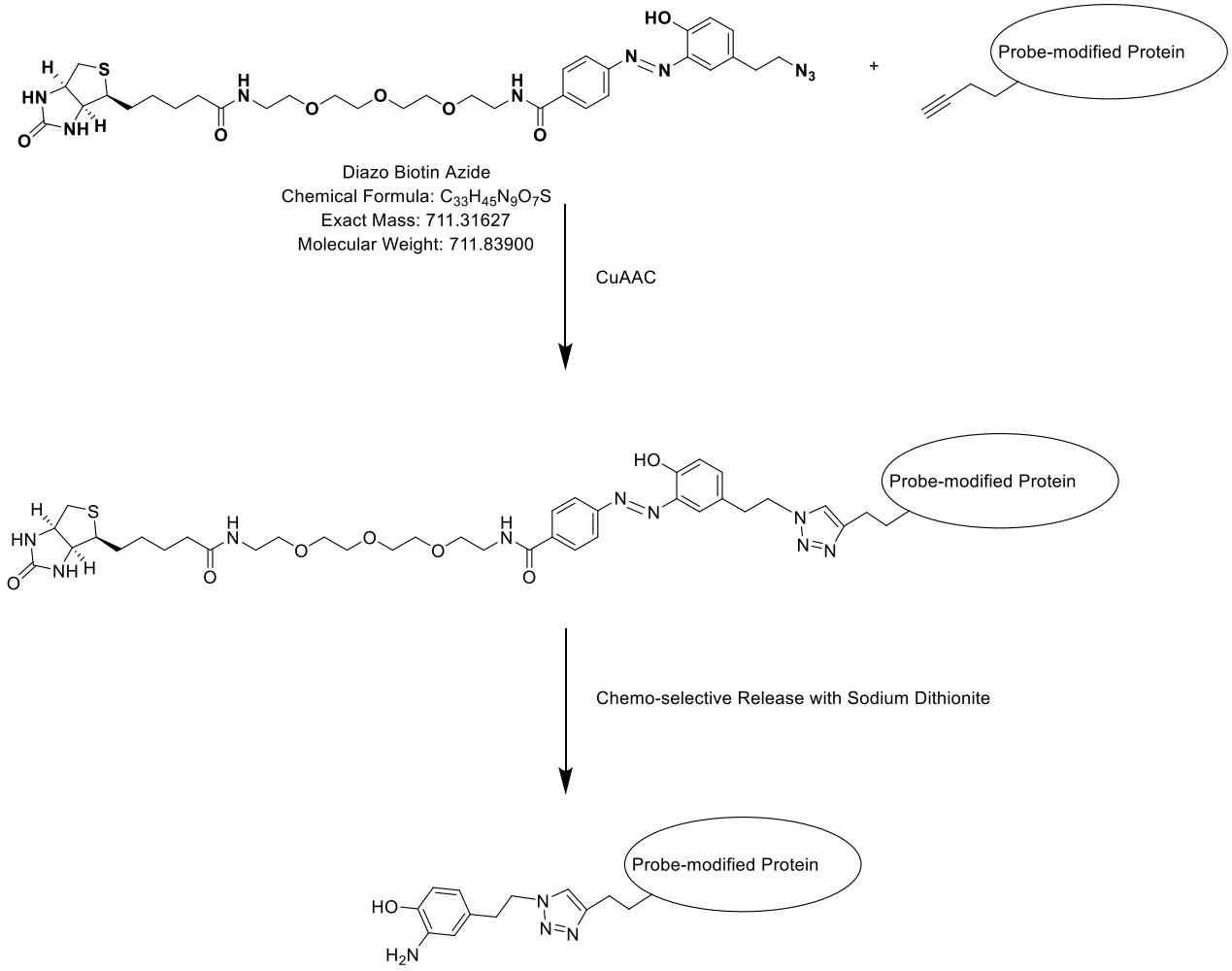
Scheme 3.5 Desthiobiotin azide tagging.



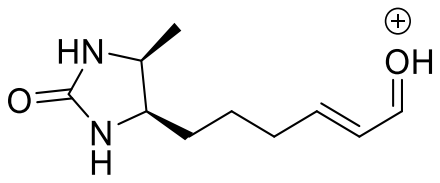
Scheme 3.6 Dde biotin picolyl azide tagging.



Scheme 3.7 Diazo biotin azide tagging.



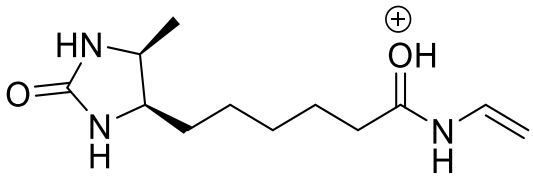
Scheme 3.8 Signature ions of Desthiobiotin-PEG$_3$.



f_1 (Des1)

Chemical Formula: C10H17N2O2+

Exact Mass: 197.1285

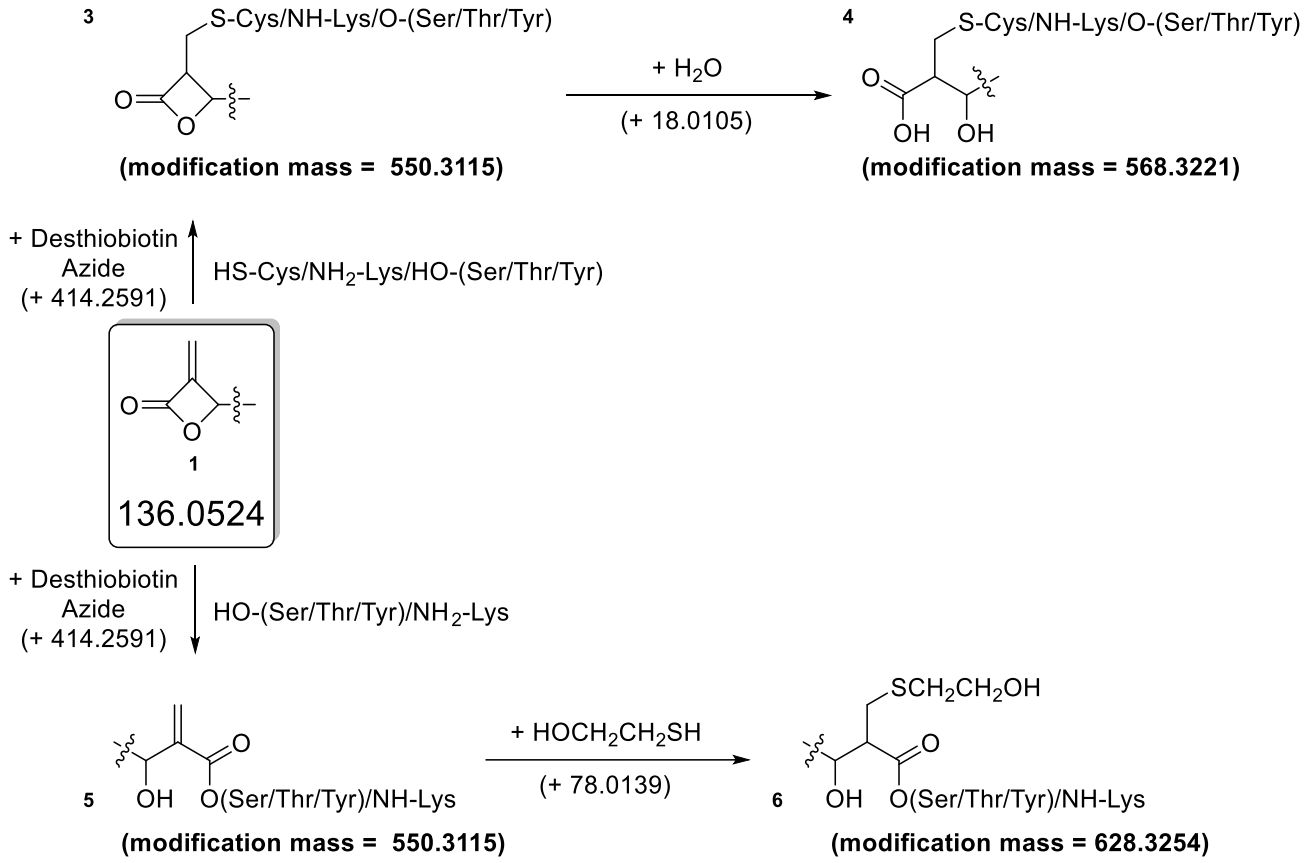


f_2 (Des2)

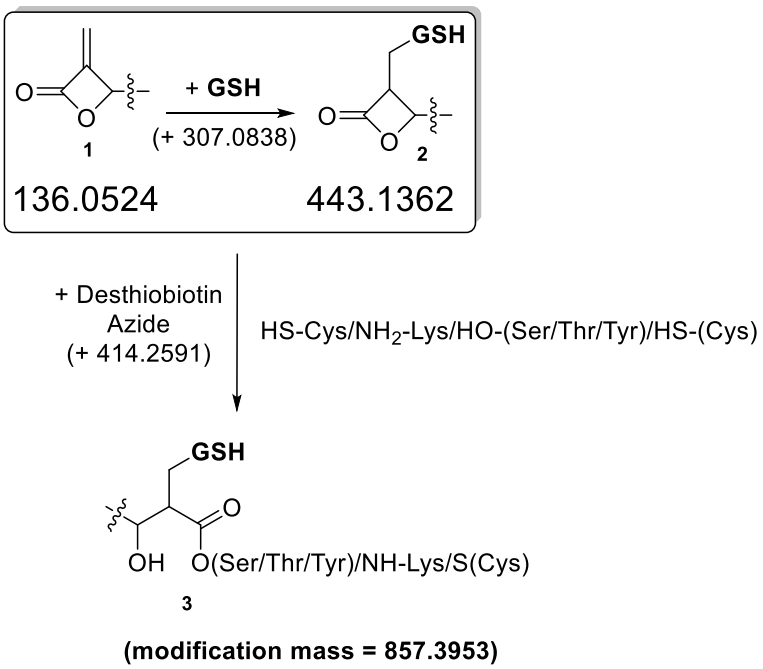
Chemical Formula: C12H22N3O2+

Exact Mass: 240.1707

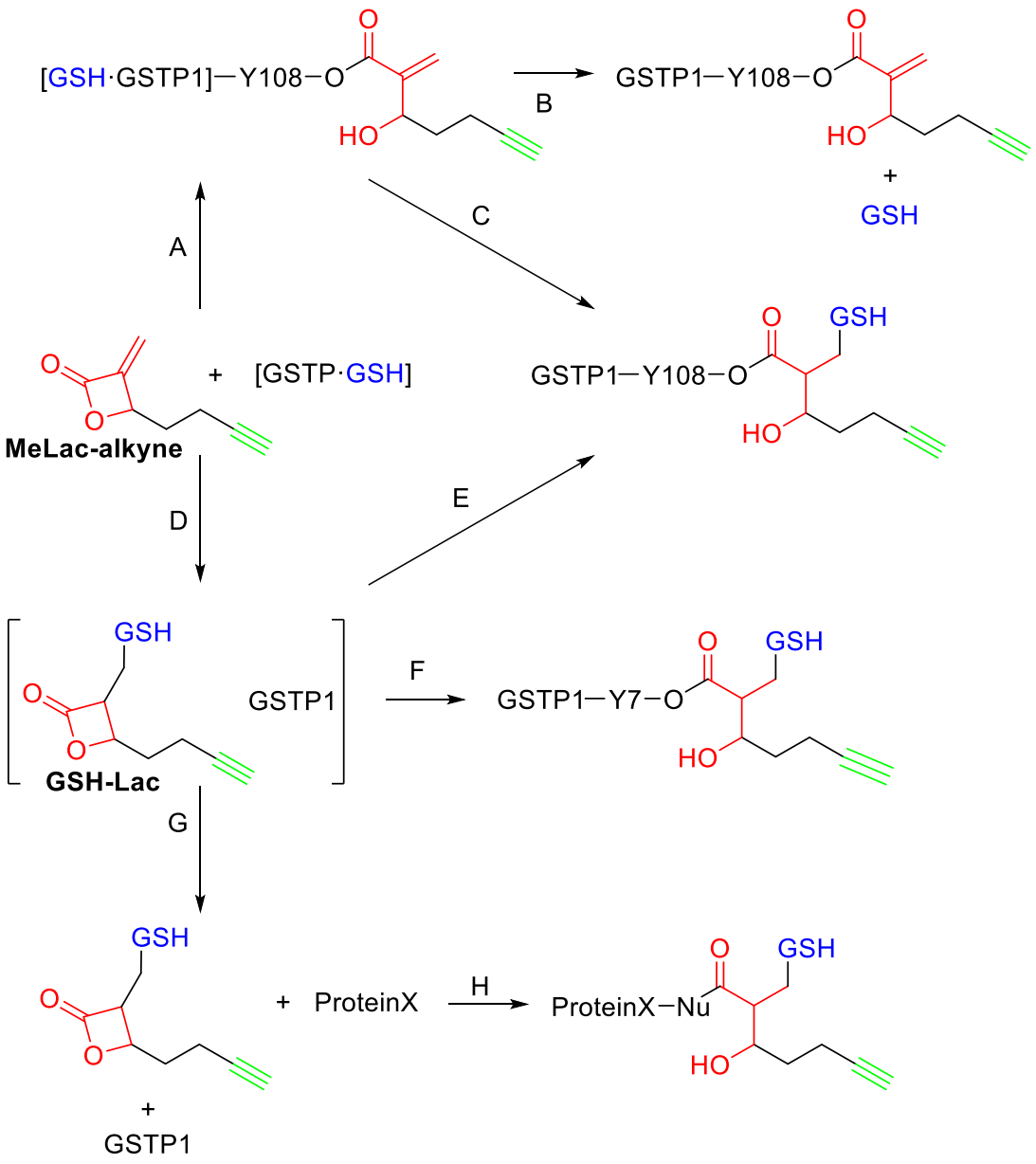
Scheme 3.9 Observed MeLac modifications by mass spectrometry.



Scheme 3.10 Observed GSH-Lac modifications by mass spectrometry.



Scheme 3.11 Proposed MeLac and GSH-Lac reaction routes with GSTP1.



3.7 Chapter 3 Tables

Table 3.1 Seeding density and confluency for various cell culture vessels.

Vessels	Surface area	Seeding density	Cells at confluency	Trypsin	Growth medium
	(cm^2)	(number of cells in millions)	(in millions)	(mL of 0.05% trypsin, 0.53 mM EDTA). Approx. volume	(mL). Approx. volume
Dishes					
35mm	8.8	0.3	1.2	1	2
60mm	21.5	0.8	3.2	3	5
100mm	56.7	2.2	8.8	5	12
150mm	145	5.0	20.0	10	30
Well plates					
6-well	9.6	0.3	1.2	1	1 to 3
12-well	3.5	0.1	0.5	0.4 to 1	1 to 2
24-well	1.9	0.05	0.24	0.2 to 0.3	0.5 to 1.0
48-well	1.1	0.03	0.12	0.1 to 0.2	0.2 to 0.4
96-well	0.32	0.01	0.04	0.05 to 0.1	0.1 to 0.2
Flasks					
T-25	25	0.7	2.8	3	3–5
T-75	75	2.1	8.4	5	8–15
T-175	175	4.9	23.3	17	35–53
T-225	225	6.3	30	22	45–68

Note: Table adapted from Ryan, John A. (2005). Growing more cells: A simple guide to small volume cell culture scale-up (Application Note No. CLS-AN-064). Retrieved from Corning Inc. website: https://www.corning.com/catalog/cls/documents/application-notes/CLS-AN-064.pdf

Table 3.2 Probe-reacted proteins with reported active sites according to M-CSA.

|Uniprot ID|Gene name	Protein name	PDB model	Known catalytical residues* (roles)	Known catalytical residues* (role type)	MaxQuant PSM score	Modification localization (Probabilities)	Modification mass (Da)	Modified residue
|P05091|ALDH2_HUMAN	Aldehyde dehydrogenase, mitochondrial	1o04	Lys192 (electrostatic stabilizer) and Cys302 (covalent catalysis, proton shuttle)	Lys192 (spectator) and Cys302 (reactant)	52.247	LLC(1)GGGIAADR	550.3115 + 18.0106	Cys
					48.513	LLC(1)GGGIAADR	550.3115	Cys
					114.58	SPNIIMSDADMDWAVEQAHFALFFNQGQC(0.085)C(0.454)C(0.461)AGSR	550.3115	Cys
					114.58	SPNIIMSDADMDWAVEQAHFALFFNQGQC(0.106)C(0.265)C(0.629)AGSR	550.3115	Cys
					89.793	SPNIIMSDADMDWAVEQAHFALFFNQGQC(0.419)C(0.419)C(0.161)AGSR	550.3115	Cys
|P13489|RINI_HUMAN	Ribonuclease inhibitor	1a4y	Lys40 (electrostatic stabilizer)	Lys40 (spectator)	61.703	ELDLSNNC(1)LGDAGILQLVESVR	550.3115 + 18.0106	Cys
					72.876	ELDLSNNC(1)LGDAGILQLVESVR	550.3115	Cys
					47.392	SNELGDVGVHC(1)VLQGLQTPSCK	550.3115 + 18.0106	Cys
					111.13	SNELGDVGVHC(1)VLQGLQTPSCK	550.3115	Cys
|P15374|UCHL3_HUMAN	Ubiquitin carboxyl-terminal hydrolase isozyme L3	1uch	Cys95 (covalent catalysis, proton shuttle)	Cys95 (reactant)	176.69	VT(0.153)HET(0.847)SAHEGQTEAPSIDEK	550.3115 + 18.0106	Thr
					56.54	VT(0.475)HET(0.524)SAHEGQT(0.001)EAPSIDEK	550.3115	Thr
					165.3	VT(0.911)HET(0.089)SAHEGQTEAPSIDEK	550.3115 + 18.0106	Thr
					100.97	VT(0.979)HET(0.021)SAHEGQTEAPSIDEK	550.3115	Thr
					176.69	VTHETS(1)AHEGQTEAPSIDEK	550.3115 + 18.0106	Ser
					84.249	VTHETS(1)AHEGQTEAPSIDEK	550.3115	Ser
|P16455|MGMT_HUMAN	Methylated-DNA--protein-cysteine methyltransferase	1eh6	Tyr114 (hydrogen bond donor, proton donor) and Cys145 (nucleophile, proton donor)	Tyr114 (interaction, reactant) and Cys145 (interaction, reactant)	60.49	GNPVPILIPC(1)HR	550.3115 + 18.0106	Cys
					65.842	GNPVPILIPC(1)HR	550.3115	Cys
					51.495	VVC(1)SSGAVGNYSGGLAVK	550.3115	Cys
|P17655|CAN2_HUMAN	Calpain-2 catalytic subunit	1kfu	Cys105 (covalently attached, proton shuttle, electrostatic stabilizer)	Cys105 (interaction, reactant, spectator)	69.594	WNDNC(1)PSWNTIDPEER	550.3115 + 18.0106	Cys
					85.77	WNDNC(1)PSWNTIDPEER	550.3115	Cys

|Q05086| UBE3A_ HUMAN	Ubiquitin-protein ligase E3A	1c4z	Cys86 (nucleofuge) and Cys 820 (covalent catalysis)	Cys86 (reactant) and Cys 820 (reactant)	69.01	GAPNNSC(1)SEIK	550.3115 + 18.0106	Cys
|Q14790| CASP8_ HUMAN	Caspase-8	1qtn	Cys360 (covalently attached)	Cys360 (intereaction)	51.875	VFFIQAC(1)QGDN YQK	550.3115	Cys
|Q93009| UBP7_H UMAN	Ubiquitin carboxyl-terminal hydrolase 7	1nbf	Cys223 (covalently attached, electrostatic stabiliser)	Cys223 (interaction, spectator)	51.092	GTC(1)VEGTIPK	550.3115 + 18.0106	Cys

Note: *Catalytical residue information was obtained from Mechanism and Catalytic Site Atlas (M-CSA)[257] data repository. According to M-CSA, a total of 667 human proteins were reported to have catalytical residues with their side chains involved. As shown above, among these 667 proteins, 8 proteins were identified as MeLac-modified ones. Among these 8 proteins, 3 proteins had MeLac modification on reported catalytical residue side chains (highlighted in red).

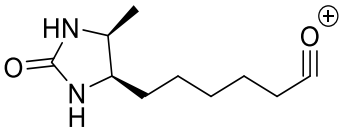
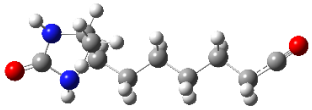
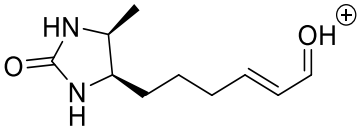
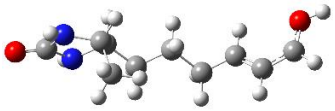
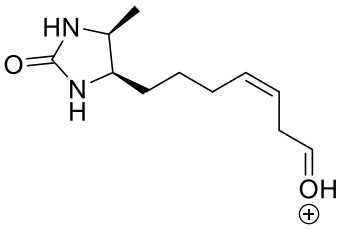
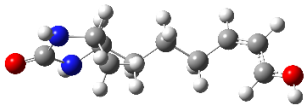
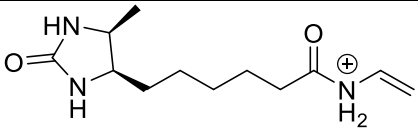
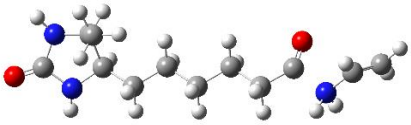
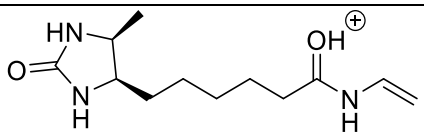
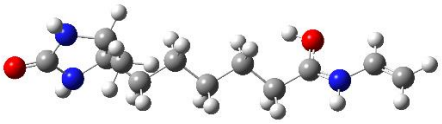
Table 3.3 Specificity of Des signature ions.

f_1			
Molecular formula	Monoisotopic mass	Unsaturation	Error (ppm)
$C_{10}H_{17}N_2O_2^+$	197.1284542	3.5	-0.2323
$C_8H_{15}N_5O^+$	197.1271115	4	-7.0434

f_2			
Molecular formula	Monoisotopic mass	Unsaturation	Error (ppm)
$C_{12}H_{22}N_3O_2^+$	240.1706534	3.5	-0.1941
$C_{14}H_{24}O_3^+$	240.1719961	3	5.3964
$C_{17}H_{22}N^+$	240.1746761	7.5	16.5552

Note:
Element composition range: C0-30, H0-60, N0-5, O0-10, S0-5.
Error range (ppm): 20.
Formula calculation used software from
http://www.cheminfo.org/Spectra/Mass/MF_from_monoisotopic_mass_and_PubChem/index.html.

Table 3.4 Geometry optimization results of Des signature ions.

Ion	2D structure	Minimized geometry	Total energy (B3LYP, Hartree)
Des1-1			-651.170362939
Des1-2 (f_1)			-651.174790289
Des1-3			-651.172576173
Des2-1			-785.206684442
Des2-2 (f_2)			-785.237450827

Note: The geometry minimization of Des signature ions was performed using Gaussian 09 (Revision D.01) at b3lyp/6-311+g(d,p) level of theory. Input ion structures were built with GaussView 5.0.9. The minimized structures (f_1 and f_2) were consistent with the previously reported fragments.[228]

Chapter 4 maxabpp as an R Package for Augmented Visualization of Peptide-centric Competitive Activity-based Protein Profiling Data from MaxQuant

4.1 Introduction

ABPP platforms are powerful chemoproteomics tools for analyzing proteome-wide perturbation introduced to living systems by foreign small molecules.[74,75] ABPP platforms are used routinely to measure the proteome-wide action of underivatized covalent drugs, environmental toxins, and reactive metabolites from the human microbiota.[76-79] When conducted for analyzing selectivity, target engagement, and off-target effects of a small molecule, an ABPP experiment usually uses competition strategy for indirect measurements of small molecule pre-occupied protein sites by comparing the probe modification difference of small molecule pre-treated sample and the control. The mass spectrometry readout of these samples can be tandem mass spectra of either probe-modified plus unmodified peptides from proteolysis of probe-labeled proteins or probe-modified peptides only.

MaxQuant is the most appreciated free software suite for analyzing MS-based proteomics profiling data[163]. The MaxQuant suite offers an easy-to-use graphic user interface (GUI), a robust peptide-spectrum matching and scoring algorithm named Andromeda,[135] and various user-configurable search parameters for tunable sensitivity and accuracy. The output of a MaxQuant run is a "txt" folder consisting of several tab-delimited tables, formatting the processed data according to different aspects of downstream analysis. These tables are intended to be read by MaxQuant and its sister software package Perseus[258] for browsing, manual validation, processing, and visualization of the database search output.

In a conventional proteomics profiling experiment, such as comparing two sets of proteomes from two different biological conditions for their qualitative and quantitative differences in expression of individual proteins, the "MaxQuant plus Perseus" workflow is usually an excellent choice for deriving the statistical inference and biological interpretation from the data. Unfortunately, this workflow does not meet some special demands in analyzing the data from various competitive ABPP experiments.[197] Without additional software support for processing ABPP data, most MaxQuant users are often forced to explore the "txt" folder manually using Microsoft Excel. These users eventually find themselves overwhelmed and frustrated by these error-prone manual steps of data manipulation.

Herein, we introduce the maxabpp package that provides a practical software tool for comprehensive downstream analysis of MaxQuant output in the R environment[259]. It is built for practicability, simplicity, flexibility, and transparency. The maxabpp package is open-source and actively maintained at https://github.com/devradiumking/maxabpp. It does not depend on any bioinformatics package, which ensures its compatibility with most user's R environments. With maxabpp and minimal efforts, novice users can perform highly reproducible and scalable data analysis that summarizes and visualizes ABPP results ready for biological interpretation. All the essential data analysis only requires a few files from the MaxQuant "txt" folder as the input, a user-created metadata.txt, and a few lines of code following the tutorial as the instruction. On the other hand, experienced R programmers can customize most functions in maxabpp to generate personalized reports. Biostatisticians and bioinformaticians can also test and implement more sophisticated statistical models on existing functions in the maxabpp package.

4.1.1 Data analysis using maxabpp

The maxabpp package contains two modules: a qualitative analysis module for cross-sample comparisons of identified proteins and a quantitative analysis module for label-free quantitation (LFQ) of probe-modified peptides.

As shown in **Figure 4.1A**, the qualitative module begins with reading the proteinGroups.txt files generated from multiple separate MaxQuant runs on the MS raw data files from samples processed in different conditions. Proteins labeled by different covalent probes or tagged by different affinity handles can be compared within and across sample groups. The proteinGroups.txt files are renamed and placed in a user-specified folder under the R working directory. Reverse sequences automatically are rejected. Proteins can be compared according to the specified column of either Protein IDs, Majority protein IDs, Protein names or Gene names on the proteinGroups.txt files as a function argument:

```
#read .txt files as datasets, for instance, condition1.txt, condition2.txt, and condition3.txt, from a folder named #“proteinGroups” under the R working directory
datasets <- read_proteinGroups(folderName = "proteinGroups")
#extract the Protein IDs column from each file and combine as setList
setList <- make_custom_setList(datasets, “Protein IDs”)
```

The maxabpp package naturally supports the conventional approach that visualizes the intersections of identified protein sets on a Venn Diagram. In addition to the Venn Diagram, maxabpp plots the Target Diagram as a novel alternative for visualizing the intersecting protein sets:

```
#generate a Venn object by calling Max_Venn function on the setList
Venn_object <- Max_Venn(setList, IndividualAnalysis = FALSE)
#plot Venn Diagram by calling plot_Max_Venn function on the Venn object
```

```
plot_Max_Venn(Venn_object)
```

As shown on **Figure 4.1B**, the quantitative module begins with reading the ModificationSpecificPeptides.txt file generated from a single MaxQuant run on multiple MS raw data files from samples treated with inhibitor(s) at various inhibitor concentrations. A user-created metadata.txt file listing all the raw data files and their corresponding experimental conditions is also required:

```
#read the .txt file as a dataset
dataset <- read_tsv("ModificationSpecificPeptides.txt")
#read the metadata.txt file
metadata <- read_tsv("metadata.txt")
#call pairwise_LFQ function to generate a LFQ table
LFQ_table <- pairwise_LFQ(
  raw = dataset,
  metadata = metadata,
  name_probe_mod = c("Mod"),
  max_each_mod = 1,
  max_total_mods = 1,
  quantitation_level = "peptide",
  background_check = FALSE,
  normalize_to = "sum_all")
#append additional information to the LFQ table
LFQ_table <- append_ec_sites(LFQ_table, quantitation_level = "peptide")
#generate the volcano plot
```

plot_volcano(LFQ_table, "InhibitorConcentration1 _vs_ InhibitorConcentration2 _log2fold_change" , " InhibitorConcentration1 _vs_ InhibitorConcentration2 _-log10p-value", xlim = c(-5.5, 2), ylim = c(0, 4), "Gene Names", pCutoff = 1.3, FCcutoff = 0, "InhibitorConcentration1 _vs_ InhibitorConcentration2/Probe Name")

Notably, the pairwise_LFQ function is highly versatile and flexible. Users can toggle parameters such as “quantitation_level”, “background_check” and “normalize_to” to perform additional analysis on the same set of data. For instance, when quantitation_level = “protein” is set, the pairwise_LFQ will perform quantitation at the protein level by aggregating/summarizing the intensity of all the probe-modified peptides belonging to individual protein groups. When the background_check = “TRUE” is set, the pairwise_LFQ will perform quantitation on background peptides (peptides without any probe modification) so that the user can assess the run-to-run variation of each sample and decide if it would be necessary to apply the normalization to the final LFQ output.

4.1.2 maxabpp Enables specialized MaxQuant data analysis for competitive ABPP platforms.

Occasionally, the cross-sample comparison of identified proteins can be a frustrating task. As part of the search algorithm, MaxQuant aggregates indistinguishable sequences, usually protein isoforms and homologous sequences, into protein groups. The aggregation of these proteins results in concatenated protein IDs, delimited by a semicolon. For example, the 5-AMP-activated protein kinase subunit gamma-1 is identified as a protein group of five isoforms/homologous sequences, "F8VYY9; A0A024R125; P54619-2; P54619; P54619-3; F8VPF5". When comparing multiple sets of protein IDs for intersections, the conventional approach simply treats each protein group as a sequence of characters (string). Consequently, when related by slightly different protein groups (scrambled or missing protein ID components) from different sets are compared, the string-matching algorithm would return FALSE, which reports these protein groups are totally different, for instance, "F8VYY9; A0A024R125; P54619-2; P54619; P54619-3; F8VPF5" vs. "P54619; A0A024R125; P54619-2; P54619-3; F8VPF5; F8VYY9" (scrambled components) vs. "F8VYY9; P54619-2; P54619; P54619-3; F8VPF5" (missing component A0A024R125) vs. "P54619-3" (single isoform). Alternatively, users often use gene names or protein names for protein IDs comparison, which drops the isoform-specific protein IDs and treats protein groups as a single canonical protein or a smaller group of homologous canonical proteins. This simplification is not ideal for processing ABPP data because protein isoform selectivity of both chemical probes and inhibitors is also considered valuable information and should be preserved. In the case of peptide-centric and site-specific analysis, distinguishable protein isoforms should be treated as individual entities rather than part of protein groups.

Therefore, a protein group-compatible alternative to Venn Diagram is needed to illustrate the intersection and exclusion relationships among various protein ID profiles obtained from

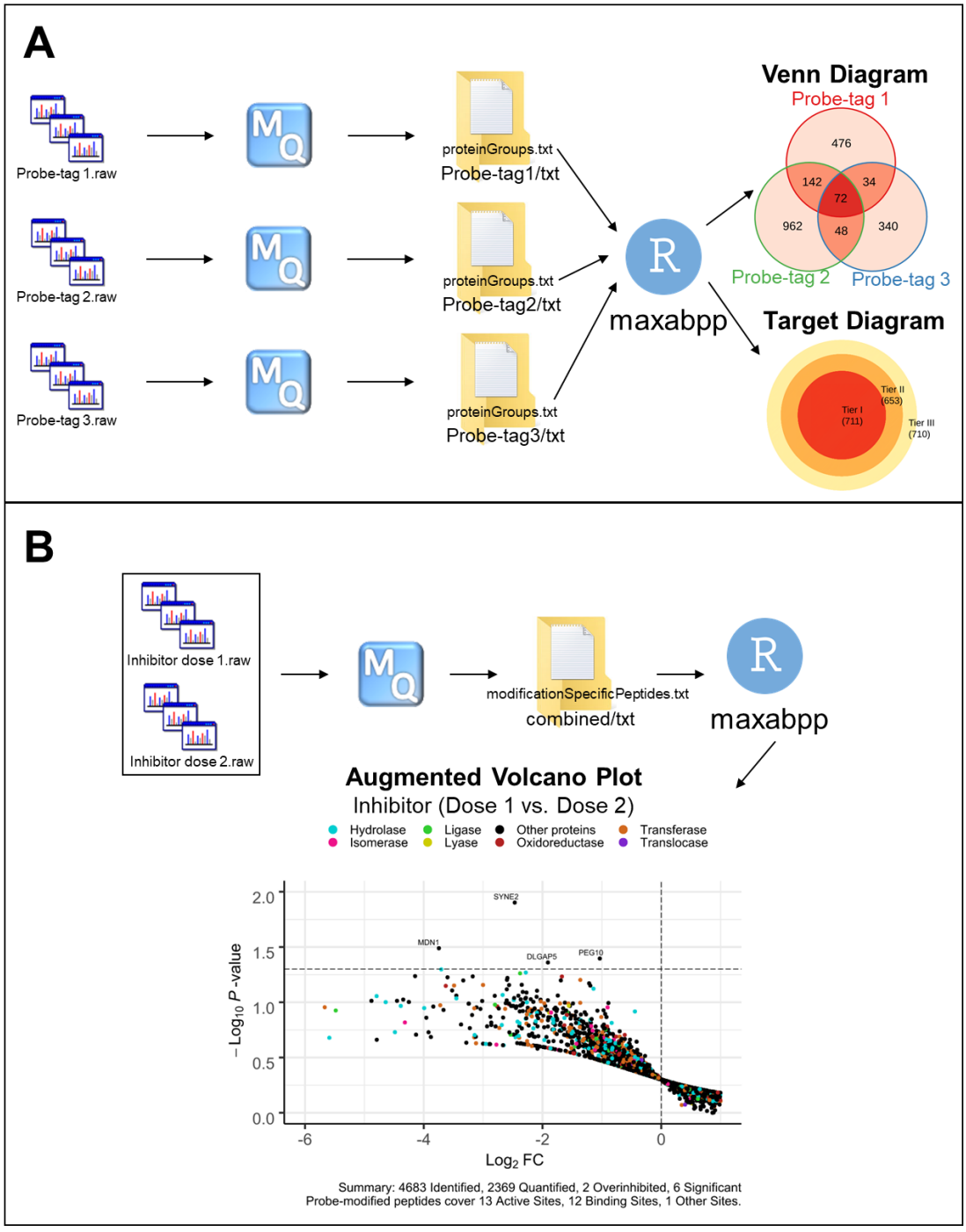
distinct sample processing methods. The qualitative module of maxabpp features the Target Diagram as a novel alternative for visualizing the intersecting protein sets. For instance, during the development of a new chemical probe, multiple distinct affinity tags may be used to explore its reactivity within a model proteome. After the CuAAC and affinity pulldown, probe-labeled proteins are enriched from each sample within a unique sample matrix. Each pulldown sample produces a unique LC-MS/MS profile. At the protein level, MaxQuant identified proteins are compared for similarities and differences. Proteins commonly identified in multiple differentially tagged samples are considered as target candidates with higher analytical confidence.[197]

On the other hand, the affinity enrichment of analytes can also be performed on probe-modified peptides after the trypsin digestion instead of probe-labeled proteins before the trypsin digestion. In this case, the maxabpp data analysis is focused on probe-modified peptides; data of probe-modified peptides needs to be filtered from the background peptides, annotated, summarized, and compared across different samples.

In summary, maxabpp is a powerful R package that integrates unique functions processing both qualitative and quantitative proteomics data, and thus provides a reliable software solution for MaxQuant downstream data analysis and visualization of competitive ABPP experiments.

4.2 Chapter 4 Figures

Figure 4.1 Overview of maxabpp workflows.



Note: (A) Qualitative workflow performing cross-sample comparisons of identified proteins. (B) Quantitative workflow performing pairwise LFQ of probe-modified peptides from competitive ABPP experiments for profiling inhibitor dose response.

Chapter 5 Conclusion and Impact

Chemical proteomics is attracting a growing amount of attention chemical biology research and beyond. Over the years, various chemical tools and analytical methods have been developed for chemical proteomics and thus significantly expediting numerous studies of small molecule modes of action and protein functions. Compared to traditional ligand binding assays, chemical proteomics methods uniquely highlight the omics methodology of holism and create systemic solutions to complex problems in chemical biology. Chemical proteomics is a multidisciplinary subject. The rapid expansion of chemical proteomics depends on its multiplying knowledge base and evolving technology, which touches a wide range of scientific subjects. The technological advancement of chemical proteomics has been discussed in three domains: chemical tools, mass spectrometry-based analytical methods, and bioinformatics support. The domain of chemical tools contains both chemical probes for protein reactivity profiling and chemical reagents for enhancing analytical performance. The former is discussed independently while the latter is considered part of the method development in mass spectrometry-based analysis.

This dissertation has presented the implementation of chemical proteomics and utilization of chemical probes for biochemical measurements in two disparate directions. The 2-nitro-ICG probe exemplifies a novel compound-centric chemical probe that answers how 2-nitroimidazole targets tumor hypoxia. This study concludes that 2-nitro-ICG and its reduced fragments modified mouse albumin as the primary protein target, but at two structurally distinct sites, possibly via two different mechanisms. The development and application of 2-nitro-ICG also demonstrates various analytical benefits, challenges, and pitfalls in the compound-centric methodological branch of chemical proteomics.

In contrast to the 2-nitro-ICG probe, the MeLac-alkyne probe is an innovative activity-based chemical probe with multiple electrophilic sites. This study of MeLac warhead and MeLac-alkyne probe explores the significance of broad reactivity in activity-based probes for their analytical versatility. The characterization of MeLac reactions concludes it is reactive to amino, hydroxyl, and thiol groups on different types of amino acid residues. It also leads to the discovery of MeLac-alkyne glutathione adduct, which reveals a potential scalable route towards rapidly conversion of a multi-electrophile activity-based probe into a large variety of compound-centric probes. These traits make the multi-electrophile a feature of greater significance on novel warheads like MeLac. Chemical probes equipped with these warheads afford distinct advantages of being both activity-based probes and scaffolds for assembling compound-centric probes simultaneously.

In addition to the exploration of chemical probes and their reactivity profiles, the development of analytical methods in mass spectrometry plays another crucial role in chemical proteomics. Mass spectrometry is intrinsically a qualitative analytical approach while biological inferences are often derived from quantitative analysis of small molecule-treated proteomes in chemical proteomics. Although label-free quantitation is a viable option in some cases of proteomics profiling, accurate quantitative mass spectrometry primarily relies on the introduction of various forms of stable isotope-based labels and internal standards. Quantitation-oriented sample preparation methodologies such as metabolic labeling on proteins, chemical tagging on peptides, and stable isotope dilution have made mass spectrometry technology amenable for accurate and precise quantitation of protein targets throughout the chemical proteomics pipeline.

Another focus in analytical method development is reduction of sample complexity. Biological samples have complex matrices. The introduction of chemical probes and click chemistry reagents to these samples further exacerbates the complexity of their matrices. Although

the affinity capture technology can significantly reduce sample complexity and efficiently enrich analytes, the final proteomics profiling data may still contain a large number of ambiguous spectra. This spectrum ambiguity issue has been demonstrated and addressed in two aspects: sample preparation and bioinformatics. In the preparation of MeLac-treated protein samples, the affinity tagging triplication technique has exposed a novel methodology for enhancing the analytical implications. On the other hand, the modification-specific data processing principle has provoked awareness and thoughts on tailoring bioinformatics tools for chemical proteomics.

Despite a few limitations, chemical proteomics has established its irreplaceable position in chemical biology. Researchers continue to exert efforts to make chemical proteomics technologies more universally applicable. The chemical tools, analytical workflows, and bioinformatics support continue to improve. As more innovations emerge at the cutting edge of chemical proteomics, we may only anticipate a faster expansion of this field.

References

1 Hood, L. & Rowen, L. The Human Genome Project: big science transforms biology and medicine. *Genome Med* **5**, 79 (2013).

2 Hutchison, C. A., 3rd. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**, 6227-6237 (2007).

3 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

4 Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L. *et al.* Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae. *Molecular & Cellular Proteomics* **1**, 323 (2002).

5 Kell, D. B. & Oliver, S. G. The metabolome 18 years on: a concept comes of age. *Metabolomics* **12**, 148 (2016).

6 Oliver, S. G., Winson, M. K., Kell, D. B. & Baganz, F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **16**, 373-378 (1998).

7 Belczacka, I., Latosinska, A., Metzger, J., Marx, D., Vlahou, A., Mischak, H. *et al.* Proteomics biomarkers for solid tumors: Current status and future prospects. *Mass Spectrometry Reviews* **38**, 49-78 (2019).

8 Isabella, V. M., Ha, B. N., Castillo, M. J., Lubkowicz, D. J., Rowe, S. E., Millet, Y. A. *et al.* Development of a synthetic live bacterial therapeutic for the human metabolic disease phenylketonuria. *Nature Biotechnology* **36**, 857-864 (2018).

9 Zetterberg, H. & Burnham, S. C. Blood-based molecular biomarkers for Alzheimer's disease. *Mol Brain* **12**, 26 (2019).

10 Frantzi, M., Bhat, A. & Latosinska, A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med* **3**, 7-7 (2014).

11 Nassar, A. F., Wu, T., Nassar, S. F. & Wisnewski, A. V. UPLC-MS for metabolomics: a giant step forward in support of pharmaceutical research. *Drug Discov Today* **22**, 463-470 (2017).

12 Schirle, M., Bantscheff, M. & Kuster, B. Mass Spectrometry-Based Proteomics in Preclinical Drug Discovery. *Chemistry & Biology* **19**, 72-84 (2012).

13 Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal Chem* **90**, 144-165 (2018).

14 Tastan, C., Karhan, E., Zhou, W., Fleming, E., Voigt, A. Y., Yao, X. *et al.* Tuning of human MAIT cell activation by commensal bacteria species and MR1-dependent T-cell presentation. *Mucosal Immunol* **11**, 1591-1605 (2018).

15 Bravo-Merodio, L., Williams, J. A., Gkoutos, G. V. & Acharjee, A. -Omics biomarker identification pipeline for translational medicine. *Journal of Translational Medicine* **17**, 155 (2019).

16 Borrebaeck, C. A. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat Rev Cancer* **17**, 199-204 (2017).

17 Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X. *et al.* Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **567**, 257-261 (2019).

18 Hedl, T. J., San Gil, R., Cheng, F., Rayner, S. L., Davidson, J. M., De Luca, A. *et al.* Proteomics Approaches for Biomarker and Drug Target Discovery in ALS and FTD. *Frontiers in Neuroscience* **13** (2019).
19 Johnson, E. C. B., Dammer, E. B., Duong, D. M., Ping, L., Zhou, M., Yin, L. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nature Medicine* **26**, 769-780 (2020).
20 Selkrig, J., Li, N., Hausmann, A., Mangan, M. S. J., Zietek, M., Mateus, A. *et al.* Spatiotemporal proteomics uncovers cathepsin-dependent macrophage cell death during Salmonella infection. *Nature Microbiology* (2020).
21 Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* (2020).
22 Lopez-Villar, E., Martos-Moreno, G. A., Chowen, J. A., Okada, S., Kopchick, J. J. & Argente, J. A proteomic approach to obesity and type 2 diabetes. *J Cell Mol Med* **19**, 1455-1470 (2015).
23 Velez, G., Roybal, C. N., Colgan, D., Tsang, S. H., Bassuk, A. G. & Mahajan, V. B. Precision Medicine: Personalized Proteomics for the Diagnosis and Treatment of Idiopathic Inflammatory Disease. *JAMA Ophthalmology* **134**, 444-448 (2016).
24 Wu, T., Ding, H., Han, J., Arriens, C., Wei, C., Han, W. *et al.* Antibody-Array-Based Proteomic Screening of Serum Markers in Systemic Lupus Erythematosus: A Discovery Study. *Journal of Proteome Research* **15**, 2102-2114 (2016).
25 Schweppe, D. K., Chavez, J. D., Lee, C. F., Caudal, A., Kruse, S. E., Stuppard, R. *et al.* Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proceedings of the National Academy of Sciences* **114**, 1732 (2017).
26 Smith, J. G. & Gerszten Robert, E. Emerging Affinity-Based Proteomic Technologies for Large-Scale Plasma Profiling in Cardiovascular Disease. *Circulation* **135**, 1651-1664 (2017).
27 McDonald, W. H. & Yates, J. R., 3rd. Shotgun proteomics: integrating technologies to answer biological questions. *Curr Opin Mol Ther* **5**, 302-309 (2003).
28 Yates, J. R., 3rd. Recent technical advances in proteomics. *F1000Res* **8** (2019).
29 Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D. *et al.* Measurement of protein using bicinchoninic acid. *Analytical Biochemistry* **150**, 76-85 (1985).
30 Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. Protein measurement with the Folin phenol reagent. *J Biol Chem* **193**, 265-275 (1951).
31 Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**, 248-254 (1976).
32 Kresge, N., Simoni, R. D. & Hill, R. L. The Discovery of Avidin by Esmond E. Snell. *Journal of Biological Chemistry* **279**, e5 (2004).
33 Diamandis, E. P. & Christopoulos, T. K. The biotin-(strept)avidin system: principles and applications in biotechnology. *Clinical Chemistry* **37**, 625-636 (1991).
34 Tsiatsiani, L. & Heck, A. J. Proteomics beyond trypsin. *FEBS J* **282**, 2612-2626 (2015).
35 Villen, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **3**, 1630-1638 (2008).

36 Sun, Z., Ji, F., Jiang, Z. & Li, L. Improving deep proteome and PTMome coverage using tandem HILIC-HPRP peptide fractionation strategy. *Anal Bioanal Chem* **411**, 459-469 (2019).
37 Batth, T. S., Francavilla, C. & Olsen, J. V. Off-Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. *Journal of Proteome Research* **13**, 6176-6186 (2014).
38 van Deemter, J. J., Zuiderweg, F. J. & Klinkenberg, A. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chemical Engineering Science* **5**, 271-289 (1956).
39 Wolters, D. A., Washburn, M. P. & Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Analytical Chemistry* **73**, 5683-5690 (2001).
40 Ho, C. S., Lam, C. W., Chan, M. H., Cheung, R. C., Law, L. K., Lit, L. C. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* **24**, 3-12 (2003).
41 Greco, V., Piras, C., Pieroni, L., Ronci, M., Putignani, L., Roncada, P. *et al.* Applications of MALDI-TOF mass spectrometry in clinical proteomics. *Expert Rev Proteomics* **15**, 683-696 (2018).
42 Wu, C. C. & MacCoss, M. J. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr Opin Mol Ther* **4**, 242-250 (2002).
43 Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., 3rd, Bairoch, A. & Bergeron, J. J. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* **7**, 681-685 (2010).
44 Hsieh, E. J., Bereman, M. S., Durand, S., Valaskovic, G. A. & MacCoss, M. J. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J Am Soc Mass Spectrom* **24**, 148-153 (2013).
45 Collins, B. C., Gillet, L. C., Rosenberger, G., Rost, H. L., Vichalkovski, A., Gstaiger, M. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat Methods* **10**, 1246-1253 (2013).
46 Rosenberger, G., Koh, C. C., Guo, T., Rost, H. L., Kouvonen, P., Collins, B. C. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* **1**, 140031 (2014).
47 Liu, Y., Buil, A., Collins, B. C., Gillet, L. C., Blum, L. C., Cheng, L. Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* **11**, 786 (2015).
48 Anjo, S. I., Santa, C. & Manadas, B. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. *Proteomics* **17** (2017).
49 Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun* **8**, 291 (2017).
50 Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Muller, M. *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**, 964-980 (2015).
51 Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat Protoc* **10**, 426-441 (2015).

52 Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258-264, 257 p following 264 (2015).
53 Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Rost, H. L. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* **34**, 1130-1136 (2016).
54 Rost, H. L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat Methods* **13**, 777-783 (2016).
55 Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation. *Molecular & Cellular Proteomics* **9**, 2252 (2010).
56 Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M. *et al.* UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom* **20**, 1989-1994 (2006).
57 Reubsaet, L., Sweredoski, M. J. & Moradian, A. Data-Independent Acquisition for the Orbitrap Q Exactive HF: A Tutorial. *J Proteome Res* **18**, 803-813 (2019).
58 Sidoli, S., Fujiwara, R. & Garcia, B. A. Multiplexed data independent acquisition (MSX-DIA) applied by high resolution mass spectrometry improves quantification quality for the analysis of histone peptides. *Proteomics* **16**, 2095-2105 (2016).
59 Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C. & Aebersold, R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol* **14**, e8126 (2018).
60 Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111 016717 (2012).
61 Rost, H. L., Malmstrom, L. & Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Mol Biol Cell* **26**, 3926-3931 (2015).
62 Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., Cheng, L. Y., Messner, S. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* **14**, 1400-1410 (2015).
63 Doerr, A. DIA mass spectrometry. *Nature Methods* **12**, 35-35 (2015).
64 Sherman, J., McKay, M. J., Ashman, K. & Molloy, M. P. How specific is my SRM?: The issue of precursor and product ion redundancy. *PROTEOMICS* **9**, 1120-1123 (2009).
65 Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* **10**, M111 011015 (2011).
66 Gallien, S., Duriez, E., Crone, C., Kellmann, M., Moehring, T. & Domon, B. Targeted Proteomic Quantification on Quadrupole-Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* **11**, 1709 (2012).
67 Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics* **11**, 1475 (2012).

68 Bourmaud, A., Gallien, S. & Domon, B. Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. *PROTEOMICS* **16**, 2146-2159 (2016).
69 Tu, C., Li, J., Shen, S., Sheng, Q., Shyr, Y. & Qu, J. Performance Investigation of Proteomic Identification by HCD/CID Fragmentations in Combination with High/Low-Resolution Detectors on a Tribrid, High-Field Orbitrap Instrument. *PLoS One* **11**, e0160160 (2016).
70 Elias, J. E. & Gygi, S. P. in *Proteome Bioinformatics* (eds Simon J. Hubbard & Andrew R. Jones) 55-71 (Humana Press, 2010).
71 An, H. & Statsyuk, A. V. An inhibitor of ubiquitin conjugation and aggresome formation. *Chem Sci* **6**, 5235-5245 (2015).
72 Eberl, H. C., Werner, T., Reinhard, F. B., Lehmann, S., Thomson, D., Chen, P. *et al.* Chemical proteomics reveals target selectivity of clinical Jak inhibitors in human primary cells. *Scientific Reports* **9**, 14159 (2019).
73 Yang, P. Y., Liu, K., Zhang, C., Chen, G. Y., Shen, Y., Ngai, M. H. *et al.* Chemical modification and organelle-specific localization of orlistat-like natural-product-based probes. *Chem Asian J* **6**, 2762-2775 (2011).
74 Moellering, R. E. & Cravatt, B. F. How chemoproteomics can enable drug discovery and development. *Chem Biol* **19**, 11-22 (2012).
75 Lanning, B. R., Whitby, L. R., Dix, M. M., Douhan, J., Gilbert, A. M., Hett, E. C. *et al.* A road map to evaluate the proteome-wide selectivity of covalent kinase inhibitors. *Nat Chem Biol* **10**, 760-767 (2014).
76 Tuley, A. & Fast, W. The Taxonomy of Covalent Inhibitors. *Biochemistry* **57**, 3326-3337 (2018).
77 Backus, K. M., Correia, B. E., Lum, K. M., Forli, S., Horning, B. D., Gonzalez-Paez, G. E. *et al.* Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534**, 570-574 (2016).
78 Hacker, S. M., Backus, K. M., Lazear, M. R., Forli, S., Correia, B. E. & Cravatt, B. F. Global profiling of lysine reactivity and ligandability in the human proteome. *Nat Chem* **9**, 1181-1190 (2017).
79 Cuesta, A. & Taunton, J. Lysine-Targeted Inhibitors and Chemoproteomic Probes. *Annu Rev Biochem* **88**, 365-381 (2019).
80 Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10335-10340 (2002).
81 Ahn, K., Johnson, D. S., Fitzgerald, L. R., Liimatta, M., Arendse, A., Stevenson, T. *et al.* Novel mechanistic class of fatty acid amide hydrolase inhibitors with remarkable selectivity. *Biochemistry* **46**, 13019-13030 (2007).
82 Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904 (2003).
83 Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169 (2004).

84 Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F. *et al.* Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Anal Chem* **80**, 2921-2931 (2008).
85 Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75**, 6843-6852 (2003).
86 Regnier, F. E. & Julka, S. Primary amine coding as a path to comparative proteomics. *Proteomics* **6**, 3968-3979 (2006).
87 Morano, C., Zhang, X. & Fricker, L. D. Multiple isotopic labels for quantitative mass spectrometry. *Anal Chem* **80**, 9298-9309 (2008).
88 Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* **4**, 484-494 (2009).
89 Dadvar, P., O'Flaherty, M., Scholten, A., Rumpel, K. & Heck, A. J. A chemical proteomics based enrichment technique targeting the interactome of the PDE5 inhibitor PF-4540124. *Mol Biosyst* **5**, 472-482 (2009).
90 Zhai, J., Liu, X., Huang, Z. & Zhu, H. RABA (reductive alkylation by acetone): a novel stable isotope labeling approach for quantitative proteomics. *J Am Soc Mass Spectrom* **20**, 1366-1377 (2009).
91 Boersema, P. J., Foong, L. Y., Ding, V. M., Lemeer, S., van Breukelen, B., Philp, R. *et al.* In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol Cell Proteomics* **9**, 84-99 (2010).
92 Oe, T., Maekawa, M., Satoh, R., Lee, S. H. & Goto, T. Combining [13C6]-phenylisothiocyanate and the Edman degradation reaction: a possible breakthrough for absolute quantitative proteomics together with protein identification. *Rapid Commun Mass Spectrom* **24**, 173-179 (2010).
93 Raijmakers, R., Dadvar, P., Pelletier, S., Gouw, J., Rumpel, K. & Heck, A. J. Target profiling of a small library of phosphodiesterase 5 (PDE5) inhibitors using chemical proteomics. *ChemMedChem* **5**, 1927-1936 (2010).
94 Zinn, N., Winter, D. & Lehmann, W. D. Recombinant isotope labeled and selenium quantified proteins for absolute protein quantification. *Anal Chem* **82**, 2334-2340 (2010).
95 Glen, A., Evans, C. A., Gan, C. S., Cross, S. S., Hamdy, F. C., Gibbins, J. *et al.* Eight-plex iTRAQ analysis of variant metastatic human prostate cancer cells identifies candidate biomarkers of progression: An exploratory study. *Prostate* **70**, 1313-1332 (2010).
96 McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C. *et al.* Increasing the Multiplexing Capacity of TMTs Using Reporter Ion Isotopologues with Isobaric Masses. *Analytical Chemistry* **84**, 7469-7478 (2012).
97 Werner, T., Becher, I., Sweetman, G., Doce, C., Savitski, M. M. & Bantscheff, M. High-Resolution Enabled TMT 8-plexing. *Analytical Chemistry* **84**, 7188-7194 (2012).
98 Thompson, A., Wölmer, N., Koncarevic, S., Selzer, S., Böhm, G., Legner, H. *et al.* TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set. *Analytical Chemistry* **91**, 15941-15950 (2019).
99 Wang, F., Chen, R., Zhu, J., Sun, D., Song, C., Wu, Y. *et al.* A fully automated system with online sample loading, isotope dimethyl labeling and multidimensional separation for high-throughput quantitative proteome analysis. *Anal Chem* **82**, 3007-3015 (2010).

100 Yao, X., Bajrami, B. & Shi, Y. Ultrathroughput multiple reaction monitoring mass spectrometry. *Anal Chem* **82**, 794-797 (2010).
101 Castillo, M. J., McShane, A. J., Cai, M., Shen, Y., Wang, L. & Yao, X. Nonisotopic reagents for a cost-effective increase in sample throughput of targeted quantitative proteomics. *Anal Chem* **87**, 9209-9216 (2015).
102 Schiess, R., Wollscheid, B. & Aebersold, R. Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol* **3**, 33-44 (2009).
103 Li, S., Diego-Limpin, P. A., Bajrami, B., Keshipeddy, S., Lam, Y.-W., Deng, B. *et al.* Scaling Proteome-Wide Reactions of Activity-Based Probes. *Analytical Chemistry* **89**, 6295-6299 (2017).
104 Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* **1**, 376-386 (2002).
105 Ong, S. E. The expanding field of SILAC. *Anal Bioanal Chem* **404**, 967-976 (2012).
106 Molina, H., Yang, Y., Ruch, T., Kim, J. W., Mortensen, P., Otto, T. *et al.* Temporal profiling of the adipocyte proteome during differentiation using a five-plex SILAC based strategy. *J Proteome Res* **8**, 48-58 (2009).
107 Zhang, G., Fenyo, D. & Neubert, T. A. Evaluation of the variation in sample preparation for comparative proteomics using stable isotope labeling by amino acids in cell culture. *J Proteome Res* **8**, 1285-1292 (2009).
108 Rangiah, K., Tippornwong, M., Sangar, V., Austin, D., Tetreault, M. P., Rustgi, A. K. *et al.* Differential secreted proteome approach in murine model for candidate biomarker discovery in colon cancer. *J Proteome Res* **8**, 5153-5164 (2009).
109 Shah, S. J., Yu, K. H., Sangar, V., Parry, S. I. & Blair, I. A. Identification and quantification of preterm birth biomarkers in human cervicovaginal fluid by liquid chromatography/tandem mass spectrometry. *J Proteome Res* **8**, 2407-2417 (2009).
110 Yu, K. H., Barry, C. G., Austin, D., Busch, C. M., Sangar, V., Rustgi, A. K. *et al.* Stable isotope dilution multidimensional liquid chromatography-tandem mass spectrometry for pancreatic cancer serum biomarker discovery. *J Proteome Res* **8**, 1565-1576 (2009).
111 Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* **7**, 383-385 (2010).
112 Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M., Ishihama, Y. *et al.* Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat Protoc* **6**, 147-157 (2011).
113 Hirsch, J. D., Eslamizar, L., Filanoski, B. J., Malekzadeh, N., Haugland, R. P., Beechem, J. M. *et al.* Easily reversible desthiobiotin binding to streptavidin, avidin, and other biotin-binding proteins: uses for protein labeling, detection, and isolation. *Analytical Biochemistry* **308**, 343-357 (2002).
114 Jung, H., Yang, T., Lasagna, M. D., Shi, J., Reinhart, G. D. & Cremer, P. S. Impact of Hapten Presentation on Antibody Binding at Lipid Membrane Interfaces. *Biophysical Journal* **94**, 3094-3103 (2008).
115 Weerapana, E., Speers, A. E. & Cravatt, B. F. Tandem orthogonal proteolysis-activity-based protein profiling (TOP-ABPP)--a general method for mapping sites of probe modification in proteomes. *Nat Protoc* **2**, 1414-1425 (2007).
116 Fukuyama, H., Ndiaye, S., Hoffmann, J., Rossier, J., Liuu, S., Vinh, J. *et al.* On-bead tryptic proteolysis: An attractive procedure for LC-MS/MS analysis of the Drosophila caspase 8

protein complex during immune response against bacteria. *Journal of Proteomics* **75**, 4610-4619 (2012).
117 Ni, X., Castanares, M., Mukherjee, A. & Lupold, S. E. Nucleic acid aptamers: clinical applications and promising new horizons. *Current medicinal chemistry* **18**, 4206-4214 (2011).
118 Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew Chem Int Ed Engl* **55**, 13408-13421 (2016).
119 Maurais, A. J. & Weerapana, E. Reactive-cysteine profiling for drug discovery. *Curr Opin Chem Biol* **50**, 29-36 (2019).
120 Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P. A. *et al.* The target landscape of clinical kinase drugs. *Science* **358** (2017).
121 Ward, C. C., Kleinman, J. I. & Nomura, D. K. NHS-Esters As Versatile Reactivity-Based Probes for Mapping Proteome-Wide Ligandable Hotspots. *ACS Chem Biol* **12**, 1478-1483 (2017).
122 Gehringer, M. & Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J Med Chem* **62**, 5673-5724 (2019).
123 Ray, S. & Murkin, A. S. New Electrophiles and Strategies for Mechanism-Based and Targeted Covalent Inhibitor Design. *Biochemistry* **58**, 5234-5244 (2019).
124 Abranyi-Balogh, P., Petri, L., Imre, T., Szijj, P., Scarpino, A., Hrast, M. *et al.* A road map for prioritizing warheads for cysteine targeting covalent inhibitors. *Eur J Med Chem* **160**, 94-107 (2018).
125 Chan, K. & o Brien, P.
126 Narayanan, A. & Jones, L. H. Sulfonyl fluorides as privileged warheads in chemical biology. *Chemical Science* **6**, 2650-2659 (2015).
127 Eng, J. K., Fischer, B., Grossmann, J. & MacCoss, M. J. A Fast SEQUEST Cross Correlation Algorithm. *Journal of Proteome Research* **7**, 4598-4602 (2008).
128 Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* **13**, 22-24 (2013).
129 McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B. *et al.* Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. *Journal of Proteome Research* **13**, 4488-4491 (2014).
130 Craig, R. & Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* **17**, 2310-2316 (2003).
131 Duncan, D. T., Craig, R. & Link, A. J. Parallel Tandem: A Program for Parallel Processing of Tandem Mass Spectra Using PVM or MPI and X!Tandem. *Journal of Proteome Research* **4**, 1842-1847 (2005).
132 Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z. & Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *PROTEOMICS – Clinical Applications* **9**, 745-754 (2015).
133 Pappin, D. J. C., Hojrup, P. & Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3**, 327-332 (1993).
134 Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5**, 5277 (2014).

135 Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **10**, 1794-1805 (2011).
136 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319 (2016).
137 Wang, L., Dietz, C., Zhou, F., Erfanzadeh, M., Zhu, Q., Smith, M. B. *et al.* Treasure hunt for peptides with undefined chemical modifications: Proteomics identification of differential albumin adducts of 2-nitroimidazole-indocyanine green in hypoxic tumor. *J Mass Spectrom* **55**, e4376 (2020).
138 Hockel, M. & Vaupel, P. Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. *J Natl Cancer Inst* **93**, 266-276 (2001).
139 Vaupel, P., Kallinowski, F. & Okunieff, P. Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: a review. *Cancer Res* **49**, 6449-6465 (1989).
140 Brahimi-Horn, M. C., Chiche, J. & Pouyssegur, J. Hypoxia and cancer. *J Mol Med (Berl)* **85**, 1301-1307 (2007).
141 Brown, J. M. & Wilson, W. R. Exploiting tumour hypoxia in cancer treatment. *Nat Rev Cancer* **4**, 437-447 (2004).
142 Vaupel, P. & Mayer, A. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev* **26**, 225-239 (2007).
143 Wilson, W. R. & Hay, M. P. Targeting hypoxia in cancer therapy. *Nat Rev Cancer* **11**, 393-410 (2011).
144 Challapalli, A., Carroll, L. & Aboagye, E. O. Molecular mechanisms of hypoxia in cancer. *Clin Transl Imaging* **5**, 225-253 (2017).
145 Biswal, N. C., Pavlik, C., Smith, M. B., Aguirre, A., Xu, Y., Zanganeh, S. *et al.* Imaging tumor hypoxia by near-infrared fluorescence tomography. *J Biomed Opt* **16**, 066009 (2011).
146 Pavlik, C., Biswal, N. C., Gaenzler, F. C., Morton, M. D., Kuhn, L. T., Claffey, K. P. *et al.* Synthesis and fluorescent characteristics of imidazole–indocyanine green conjugates. *Dyes and Pigments* **89**, 9-15 (2011).
147 Mohammad, I., Stanford, C., Morton, M. D., Zhu, Q. & Smith, M. B. Structurally modified indocyanine green dyes. Modification of the polyene linker. *Dyes and Pigments* **99**, 275-283 (2013).
148 Xu, Y., Zanganeh, S., Mohammad, I., Aguirre, A., Wang, T., Yang, Y. *et al.* Targeting tumor hypoxia with 2-nitroimidazole-indocyanine green dye conjugates. *J Biomed Opt* **18**, 66009 (2013).
149 Zanganeh, S., Li, H., Kumavor, P. D., Alqasemi, U., Aguirre, A., Mohammad, I. *et al.* Photoacoustic imaging enhanced by indocyanine green-conjugated single-wall carbon nanotubes. *J Biomed Opt* **18**, 096006 (2013).
150 Zhou, F., Zanganeh, S., Mohammad, I., Dietz, C., Abuteen, A., Smith, M. B. *et al.* Targeting tumor hypoxia: a third generation 2-nitroimidazole-indocyanine dye-conjugate with improved fluorescent yield. *Org Biomol Chem* **13**, 11220-11227 (2015).
151 Abuteen, A., Zhou, F., Dietz, C., Mohammad, I., Smith, M. & Zhu, Q. *Synthesis of a 4-Nitroimidazole Indocyanine Dye–Conjugate and Imaging of Tumor Hypoxia in BALB/c Tumor-Bearing Female Mice*. Vol. 126 (2015).

152 Hodgkiss, R. J. Use of 2-nitroimidazoles as bioreductive markers for tumour hypoxia. *Anti-cancer drug design* **13**, 687-702 (1998).
153 Kizaka-Kondoh, S. & Konse-Nagasawa, H. Significance of nitroimidazole compounds and hypoxia-inducible factor-1 for imaging tumor hypoxia. *Cancer Sci* **100**, 1366-1373 (2009).
154 Okuda, K., Okabe, Y., Kadonosono, T., Ueno, T., Youssif, B. G., Kizaka-Kondoh, S. *et al.* 2-Nitroimidazole-tricarbocyanine conjugate as a near-infrared fluorescent probe for in vivo imaging of tumor hypoxia. *Bioconjug Chem* **23**, 324-329 (2012).
155 Raleigh, J. A. & Liu, S. F. Reductive fragmentation of 2-nitroimidazoles in the presence of nitroreductases—glyoxal formation from misonidazole. *Biochemical Pharmacology* **32**, 1444-1446 (1983).
156 Raleigh, J. A. & Koch, C. J. Importance of thiols in the reductive binding of 2-nitroimidazoles to macromolecules. *Biochem Pharmacol* **40**, 2457-2464 (1990).
157 Trochine, A., Creek, D. J., Faral-Tello, P., Barrett, M. P. & Robello, C. Benznidazole biotransformation and multiple targets in Trypanosoma cruzi revealed by metabolomics. *PLoS Negl Trop Dis* **8**, e2844 (2014).
158 Masaki, Y., Shimizu, Y., Yoshioka, T., Tanaka, Y., Nishijima, K., Zhao, S. *et al.* The accumulation mechanism of the hypoxia imaging probe "FMISO" by imaging mass spectrometry: possible involvement of low-molecular metabolites. *Sci Rep* **5**, 16802 (2015).
159 Masaki, Y., Shimizu, Y., Yoshioka, T., Feng, F., Zhao, S., Higashino, K. *et al.* Imaging Mass Spectrometry Revealed the Accumulation Characteristics of the 2-Nitroimidazole-Based Agent "Pimonidazole" in Hypoxia. *PLoS One* **11**, e0161639 (2016).
160 Mascini, N. E., Cheng, M., Jiang, L., Rizwan, A., Podmore, H., Bhandari, D. R. *et al.* Mass Spectrometry Imaging of the Hypoxia Marker Pimonidazole in a Breast Tumor Model. *Anal Chem* **88**, 3107-3114 (2016).
161 Arteel, G. E., Thurman, R. G. & Raleigh, J. A. Reductive metabolism of the hypoxia marker pimonidazole is regulated by oxygen tension independent of the pyridine nucleotide redox state. *Eur J Biochem* **253**, 743-750 (1998).
162 Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **4**, 787-797 (2007).
163 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).
164 Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology* **33**, 743 (2015).
165 Ballard, T. E., Dahal, U. P., Bessire, A. J., Schneider, R. P., Geoghegan, K. F. & Vaz, A. D. A tag-free collisionally induced fragmentation approach to detect drug-adducted proteins by mass spectrometry. *Rapid Commun Mass Spectrom* **29**, 2175-2183 (2015).
166 Dorri, Y. in *Protein Electrophoresis: Methods and Protocols* (eds Biji T. Kurien & R. Hal Scofield) 235-246 (Humana Press, 2012).
167 Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918-920 (2012).
168 UniProt, C. The universal protein resource (UniProt). *Nucleic Acids Res* **36**, D190-195 (2008).

169 Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303 (2018).
170 Majorek, K. A., Porebski, P. J., Dayal, A., Zimmerman, M. D., Jablonska, K., Stewart, A. J. *et al.* Structural and immunologic characterization of bovine, horse, and rabbit serum albumins. *Mol Immunol* **52**, 174-182 (2012).
171 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455-461 (2010).
172 Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785-2791 (2009).
173 Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E. & Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform* **4**, 17 (2012).
174 Cravatt, B. F., Wright, A. T. & Kozarich, J. W. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu Rev Biochem* **77**, 383-414 (2008).
175 Diamandis, E. P. & Christopoulos, T. K. The biotin-(strept)avidin system: principles and applications in biotechnology. *Clin Chem* **37**, 625-636 (1991).
176 Desmettre, T., Devoisselle, J. M. & Mordon, S. Fluorescence Properties and Metabolic Features of Indocyanine Green (ICG) as Related to Angiography. *Survey of Ophthalmology* **45**, 15-27 (2000).
177 Hamann, F. M., Brehm, R., Pauli, J., Grabolle, M., Frank, W., Kaiser, W. A. *et al.* Controlled modulation of serum protein binding and biodistribution of asymmetric cyanine dyes by variation of the number of sulfonate groups. *Mol Imaging* **10**, 258-269 (2011).
178 Williams, C. F., Lloyd, D., Kolarich, D., Alagesan, K., Duchene, M., Cable, J. *et al.* Disrupted intracellular redox balance of the diplomonad fish parasite Spironucleus vortens by 5-nitroimidazoles and garlic-derived compounds. *Vet Parasitol* **190**, 62-73 (2012).
179 Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics* **11**, M111 010199 (2012).
180 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
181 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214 (2007).
182 Mitrofanov, E., Muskat, T. & Grotemeyer, J. Indocyanine green MS/MS investigations using femtosecond laser-pulse photodissociation and collision-induced dissociation. *Eur J Mass Spectrom (Chichester)* **24**, 299-312 (2018).
183 Ang, C. W., Jarrad, A. M., Cooper, M. A. & Blaskovich, M. A. T. Nitroimidazoles: Molecular Fireworks That Combat a Broad Spectrum of Infectious Diseases. *J Med Chem* **60**, 7636-7657 (2017).
184 Penha, F. M., Rodrigues, E. B., Maia, M., Meyer, C. H., Costa Ede, P., Dib, E. *et al.* Biochemical analysis and decomposition products of indocyanine green in relation to solvents, dye concentrations and laser exposure. *Ophthalmologica* **230 Suppl 2**, 59-67 (2013).
185 Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics* **9**, 2252-2261 (2010).

186 Turriziani, B., Garcia-Munoz, A., Pilkington, R., Raso, C., Kolch, W. & von Kriegsheim, A. On-beads digestion in conjunction with data-dependent mass spectrometry: a shortcut to quantitative and dynamic interaction proteomics. *Biology (Basel)* **3**, 320-332 (2014).
187 Curry, S., Mandelkow, H., Brick, P. & Franks, N. Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites. *Nat Struct Biol* **5**, 827-835 (1998).
188 Kragh-Hansen, U., Chuang, V. T. & Otagiri, M. Practical aspects of the ligand-binding and enzymatic properties of human serum albumin. *Biol Pharm Bull* **25**, 695-704 (2002).
189 Hobert, E. M., Doerner, A. E., Walker, A. S. & Schepartz, A. Effective molarity redux: Proximity as a guiding force in chemistry and biology. *Isr J Chem* **53**, 567-576 (2013).
190 Ursini, F., Maiorino, M. & Forman, H. J. Redox homeostasis: The Golden Mean of healthy living. *Redox Biol* **8**, 205-215 (2016).
191 Peters, T. in *All About Albumin: Biochemistry, Genetics, and Medical Applications* (ed Theodore Peters) Ch. 5, 230-233 (Academic Press, 1995).
192 Monks, J. & Neville, M. C. Albumin transcytosis across the epithelium of the lactating mouse mammary gland. *J Physiol* **560**, 267-280 (2004).
193 Shamay, A., Homans, R., Fuerman, Y., Levin, I., Barash, H., Silanikove, N. *et al.* Expression of albumin in nonhepatic tissues and its synthesis by the bovine mammary gland. *J Dairy Sci* **88**, 569-576 (2005).
194 Peters, T. in *All About Albumin: Biochemistry, Genetics, and Medical Applications* (ed Theodore Peters) Ch. 5, 233-233 (Academic Press, 1995).
195 Soreide, J. A., Lea, O. A. & Kvinnsland, S. Cytosol albumin content in operable breast cancer. Correlations to steroid hormone receptors, other prognostic factors and prognosis. *Acta Oncol* **30**, 797-802 (1991).
196 Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**, 135-145 (2018).
197 Wang, L., Riel, L., Bajrami, B., Deng, B., Howell, A. & Yao, X. *α-Methylene-β-Lactone Probe for Measuring Live-Cell Reactions of Small Molecules*. (2020).
198 Swinney, D. C. Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines. *Clinical Pharmacology & Therapeutics* **93**, 299-301 (2013).
199 Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: origins and evolution. *Nature Reviews Drug Discovery* **13**, 577-587 (2014).
200 Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery* **16**, 531-543 (2017).
201 Gerry, C. J. & Schreiber, S. L. Chemical probes and drug leads from advances in synthetic planning and methodology. *Nature Reviews Drug Discovery* **17**, 333-352 (2018).
202 Bar-Peled, L., Kemper, E. K., Suciu, R. M., Vinogradova, E. V., Backus, K. M., Horning, B. D. *et al.* Chemical Proteomics Identifies Druggable Vulnerabilities in a Genetically Defined Cancer. *Cell* **171**, 696-709 e623 (2017).
203 Gu, C., Shannon, D. A., Colby, T., Wang, Z., Shabab, M., Kumari, S. *et al.* Chemical proteomics with sulfonyl fluoride probes reveals selective labeling of functional tyrosines in glutathione transferases. *Chem Biol* **20**, 541-548 (2013).
204 Baggelaar, M. P., Chameau, P. J., Kantae, V., Hummel, J., Hsu, K. L., Janssen, F. *et al.* Highly Selective, Reversible Inhibitor Identified by Comparative Chemoproteomics

Modulates Diacylglycerol Lipase Activity in Neurons. *J Am Chem Soc* **137**, 8851-8857 (2015).
205 Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: The serine hydrolases. *Proceedings of the National Academy of Sciences* **96**, 14694-14699 (1999).
206 Gehringer, M. & Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J Med Chem*, acs.jmedchem.8b01153-acs.jmedchem.01158b01153 (2019).
207 Lonsdale, R., Burgess, J., Colclough, N., Davies, N. L., Lenz, E. M., Orton, A. L. *et al.* Expanding the Armory: Predicting and Tuning Covalent Warhead Reactivity. *J Chem Inf Model* **57**, 3124-3137 (2017).
208 Shindo, N., Fuchida, H., Sato, M., Watari, K., Shibata, T., Kuwata, K. *et al.* Selective and reversible modification of kinase cysteines with chlorofluoroacetamides. *Nat Chem Biol* **15**, 250-258 (2019).
209 Liu, Q., Sabnis, Y., Zhao, Z., Zhang, T., Buhrlage, S. J., Jones, L. H. *et al.* Developing irreversible inhibitors of the protein kinase cysteinome. *Chem Biol* **20**, 146-159 (2013).
210 Simon, G. M. & Cravatt, B. F. Activity-based proteomics of enzyme superfamilies: serine hydrolases as a case study. *J Biol Chem* **285**, 11051-11055 (2010).
211 Parsons, W. H., Kolar, M. J., Kamat, S. S., Cognetta, A. B., 3rd, Hulce, J. J., Saez, E. *et al.* AIG1 and ADTRP are atypical integral membrane hydrolases that degrade bioactive FAHFAs. *Nat Chem Biol* **12**, 367-372 (2016).
212 Gambini, L., Baggio, C., Udompholkul, P., Jossart, J., Salem, A. F., Perry, J. J. P. *et al.* Covalent Inhibitors of Protein-Protein Interactions Targeting Lysine, Tyrosine, or Histidine Residues. *J Med Chem* **62**, 5616-5627 (2019).
213 Pettinger, J., Jones, K. & Cheeseman, M. D. Lysine-Targeting Covalent Inhibitors. *Angew Chem Int Ed Engl* **56**, 15200-15209 (2017).
214 Jones, L. H. Reactive Chemical Probes: Beyond the Kinase Cysteinome. *Angew Chem Int Ed Engl* **57**, 9220-9223 (2018).
215 Mukherjee, H., Debreczeni, J., Breed, J., Tentarelli, S., Aquila, B., Dowling, J. E. *et al.* A study of the reactivity of S((VI))-F containing warheads with nucleophilic amino-acid side chains under physiological conditions. *Org Biomol Chem* **15**, 9685-9695 (2017).
216 Ma, N., Hu, J., Zhang, Z. M., Liu, W., Huang, M., Fan, Y. *et al.* 2H-Azirine-Based Reagents for Chemoselective Bioconjugation at Carboxyl Residues Inside Live Cells. *J Am Chem Soc* (2020).
217 Kathman, S. G., Xu, Z. & Statsyuk, A. V. A fragment-based method to discover irreversible covalent inhibitors of cysteine proteases. *J Med Chem* **57**, 4969-4974 (2014).
218 Ettari, R., Micale, N., Schirmeister, T., Gelhaus, C., Leippe, M., Nizi, E. *et al.* Novel peptidomimetics containing a vinyl ester moiety as highly potent and selective falcipain-2 inhibitors. *J Med Chem* **52**, 2157-2160 (2009).
219 Böttcher, T. & Sieber, S. A. β-Lactams and β-lactones as activity-based probes in chemical biology. *MedChemComm* **3**, 408-408 (2012).
220 Yang, P. Y., Liu, K., Ngai, M. H., Lear, M. J., Wenk, M. R. & Yao, S. Q. Activity-based proteome profiling of potential cellular targets of Orlistat--an FDA-approved drug with anti-tumor activities. *J Am Chem Soc* **132**, 656-666 (2010).
221 Groll, M., Korotkov, V. S., Huber, E. M., de Meijere, A. & Ludwig, A. A Minimal beta-Lactone Fragment for Selective beta5c or beta5i Proteasome Inhibitors. *Angew Chem Int Ed Engl* **54**, 7810-7814 (2015).

222 Camara, K., Kamat, S. S., Lasota, C. C., Cravatt, B. F. & Howell, A. R. Combining cross-metathesis and activity-based protein profiling: new beta-lactone motifs for targeting serine hydrolases. *Bioorg Med Chem Lett* **25**, 317-321 (2015).
223 Malapit, C. A., Caldwell, D. R., Sassu, N., Milbin, S. & Howell, A. R. Pd-Catalyzed Acyl C-O Bond Activation for Selective Ring-Opening of alpha-Methylene-beta-lactones with Amines. *Org Lett* **19**, 1966-1969 (2017).
224 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520 (2017).
225 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319 (2016).
226 Avtonomov, D. M., Kong, A. & Nesvizhskii, A. I. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. *J Proteome Res* **18**, 715-720 (2019).
227 Hahm, H. S., Toroitich, E. K., Borne, A. L., Brulet, J. W., Libby, A. H., Yuan, K. *et al.* Global targeting of functional tyrosines using sulfur-triazole exchange chemistry. *Nature Chemical Biology* (2019).
228 Browne, C. M., Jiang, B., Ficarro, S. B., Doctor, Z. M., Johnson, J. L., Card, J. D. *et al.* A Chemoproteomic Strategy for Direct and Proteome-Wide Covalent Inhibitor Target-Site Identification. *J Am Chem Soc* **141**, 191-203 (2019).
229 Brademan, D. R., Riley, N. M., Kwiecien, N. W. & Coon, J. J. Interactive Peptide Spectral Annotator: A Versatile Web-based Tool for Proteomic Applications. *Mol Cell Proteomics* **18**, S193-S201 (2019).
230 Smith, J. M., Jami Alahmadi, Y. & Rowley, C. N. Range-Separated DFT Functionals are Necessary to Model Thio-Michael Additions. *J Chem Theory Comput* **9**, 4860-4865 (2013).
231 Harshbarger, W., Gondi, S., Ficarro, S. B., Hunter, J., Udayakumar, D., Gurbani, D. *et al.* Structural and Biochemical Analyses Reveal the Mechanism of Glutathione S-Transferase Pi 1 Inhibition by the Anti-cancer Compound Piperlongumine. *J Biol Chem* **292**, 112-120 (2017).
232 Dongré, A. R., Jones, J. L., Somogyi, Á. & Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *Journal of the American Chemical Society* **118**, 8365-8374 (1996).
233 Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).
234 Pettinger, J., Le Bihan, Y. V., Widya, M., van Montfort, R. L., Jones, K. & Cheeseman, M. D. An Irreversible Inhibitor of HSP72 that Unexpectedly Targets Lysine-56. *Angew Chem Int Ed Engl* **56**, 3536-3540 (2017).
235 Groll, M., Schellenberg, B., Bachmann, A. S., Archer, C. R., Huber, R., Powell, T. K. *et al.* A plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism. *Nature* **452**, 755-758 (2008).
236 Lin, D., Saleh, S. & Liebler, D. C. Reversibility of covalent electrophile-protein adducts and chemical toxicity. *Chem Res Toxicol* **21**, 2361-2369 (2008).

237 Abo, M. & Weerapana, E. A Caged Electrophilic Probe for Global Analysis of Cysteine Reactivity in Living Cells. *J Am Chem Soc* **137**, 7087-7090 (2015).
238 Kridel, S. J., Axelrod, F., Rozenkrantz, N. & Smith, J. W. Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity. *Cancer Res* **64**, 2070-2075 (2004).
239 Xiao, D., Shi, D., Yang, D., Barthel, B., Koch, T. H. & Yan, B. Carboxylesterase-2 is a highly sensitive target of the antiobesity agent orlistat with profound implications in the activation of anticancer prodrugs. *Biochem Pharmacol* **85**, 439-447 (2013).
240 Berdan, C. A., Ho, R., Lehtola, H. S., To, M., Hu, X., Huffman, T. R. *et al.* Parthenolide Covalently Targets and Inhibits Focal Adhesion Kinase in Breast Cancer Cells. *Cell Chem Biol* **26**, 1027-1035 e1022 (2019).
241 Freund, R. R. A., Gobrecht, P., Fischer, D. & Arndt, H. D. Advances in chemistry and bioactivity of parthenolide. *Nat Prod Rep* **37**, 541-565 (2020).
242 Dey, S., Sarkar, M. & Giri, B. Anti-inflammatory and Anti-tumor Activities of Parthenolide: An Update. *Journal of Chemical Biology & Therapeutics* **01**, 1-6 (2016).
243 Nomura, D. K. & Maimone, T. J. Target Identification of Bioactive Covalently Acting Natural Products. *Curr Top Microbiol Immunol* **420**, 351-374 (2019).
244 Wong, C. Y. Y., Choi, A. W.-T., Lui, M. Y., Fridrich, B., Horváth, A. K., Mika, L. T. *et al.* Stability of gamma-valerolactone under neutral, acidic, and basic conditions. *Structural Chemistry* **28**, 423-429 (2017).
245 Zimniak, P., Nanduri, B., Pikula, S., Bandorowicz-Pikula, J., Singhal, S. S., Srivastava, S. K. *et al.* Naturally occurring human glutathione S-transferase GSTP1-1 isoforms with isoleucine and valine in position 104 differ in enzymic properties. *Eur J Biochem* **224**, 893-899 (1994).
246 Ji, X., Tordova, M., O'Donnell, R., Parsons, J. F., Hayden, J. B., Gilliland, G. L. *et al.* Structure and function of the xenobiotic substrate-binding site and location of a potential non-substrate-binding site in a class pi glutathione S-transferase. *Biochemistry* **36**, 9690-9702 (1997).
247 Townsend, D. M. & Tew, K. D. The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* **22**, 7369-7375 (2003).
248 Dourado, D. F., Fernandes, P. A., Ramos, M. J. & Mannervik, B. Mechanism of glutathione transferase P1-1-catalyzed activation of the prodrug canfosfamide (TLK286, TELCYTA). *Biochemistry* **52**, 8069-8078 (2013).
249 Findlay, V. J., Townsend, D. M., Saavedra, J. E., Buzard, G. S., Citro, M. L., Keefer, L. K. *et al.* Tumor cell responses to a novel glutathione S-transferase-activated nitric oxide-releasing prodrug. *Mol Pharmacol* **65**, 1070-1079 (2004).
250 Tamura, T., Ueda, T., Goto, T., Tsukidate, T., Shapira, Y., Nishikawa, Y. *et al.* Rapid labelling and covalent inhibition of intracellular native proteins using ligand-directed N-acyl-N-alkyl sulfonamide. *Nat Commun* **9**, 1870 (2018).
251 Zhao, Q., Ouyang, X., Wan, X., Gajiwala, K. S., Kath, J. C., Jones, L. H. *et al.* Broad-Spectrum Kinase Profiling in Live Cells with Lysine-Targeted Sulfonyl Fluoride Probes. *J Am Chem Soc* **139**, 680-685 (2017).
252 Perez-Miller, S. J. & Hurley, T. D. Coenzyme Isomerization Is Integral to Catalysis in Aldehyde Dehydrogenase. *Biochemistry* **42**, 7100-7109 (2003).
253 Daniels, D. S., Mol, C. D., Arvai, A. S., Kanugula, S., Pegg, A. E. & Tainer, J. A. Active and alkylated human AGT structures: a novel zinc site, inhibitor and extrahelical base binding. *The EMBO Journal* **19**, 1719-1730 (2000).

254 Watt, W., Koeplinger, K. A., Mildner, A. M., Heinrikson, R. L., Tomasselli, A. G. & Watenpaugh, K. D. The atomic-resolution structure of human caspase-8, a key activator of apoptosis. *Structure* **7**, 1135-1143 (1999).
255 Strobl, S., Fernandez-Catalan, C., Braun, M., Huber, R., Masumoto, H., Nakagawa, K. *et al.* The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium. *Proceedings of the National Academy of Sciences* **97**, 588-592 (2000).
256 Papageorgiou, A. C., Shapiro, R. & Acharya, K. R. Molecular recognition of human angiogenin by placental ribonuclease inhibitor—an X-ray crystallographic study at 2.0 Å resolution. *The EMBO Journal* **16**, 5162-5177 (1997).
257 Ribeiro, A. J. M., Holliday, G. L., Furnham, N., Tyzack, J. D., Ferris, K. & Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* **46**, D618-D623 (2018).
258 Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**, 731-740 (2016).
259 Team, R. C. R: A language and environment for statistical computing.